



**MACHINE LEARNING BASED EARLY-SEASON CROP
YIELD FORECASTING USING SENTINEL-2 TIME
SERIES: AN ANALYSIS OF PREDICTION TIMING
AND ACCURACY TRADE-OFFS**

METİN GÜNEY ÖZTÜRK

Thesis for the Master's Program in Electrical and Electronics Engineering

Graduate School

Izmir University of Economics

İzmir

2026

**MACHINE LEARNING BASED EARLY-SEASON CROP
YIELD FORECASTING USING SENTINEL-2 TIME
SERIES: AN ANALYSIS OF PREDICTION TIMING
AND ACCURACY TRADE-OFFS**

METİN GÜNEY ÖZTÜRK

THESIS ADVISOR: ASST. PROF. DR. AYÇA KUMLUCA TOPALLI

Master's Exam Jury Members

Assoc. Prof. Dr. Nalan ÖZKURT

Asst. Prof. Dr. Ayça Kumluca TOPALLI

Asst. Prof. Dr. Faezeh YEGANLI

A Master's Thesis

Submitted to

the Graduate School of Izmir University of Economics
the Department of Electrical and Electronics Engineering

İzmir

2026

Approval of the Graduate School

Prof. Dr. Muhittin Hakan DEMİR

I certify that this thesis satisfies all the requirements as a thesis for a Master's degree.

Prof. Dr. Aydın AKAN

This is to certify that we have read this thesis and that in our opinion it is fully adequate, in scope and quality, as a thesis for a Master's degree.

Asst. Prof. Dr. Ayça KÜMLUCA TOPALLI

Master's Exam Jury Members

Assoc. Prof. Dr. Nalan ÖZKURT
Yaşar University

Asst. Prof. Dr. Ayça KÜMLUCA TOPALLI
İzmir University of Economics

Asst. Prof. Dr. Faezeh YEGANLI
İzmir University of Economics

ETHICAL DECLARATION

I hereby declare that I am the sole author of this thesis and that I have conducted my work in accordance with academic rules and ethical behaviour at every stage from the planning of the thesis to its defence. I confirm that I have cited all ideas, information and findings that are not specific to my study, as required by the code of ethical behaviour, and that all statements not cited are my own.

Name, Surname: Metin Güney ÖZTÜRK

Date: 16.01.2026

Signature:

ABSTRACT

MACHINE LEARNING BASED EARLY-SEASON CROP YIELD FORECASTING USING SENTINEL-2 TIME SERIES: AN ANALYSIS OF PREDICTION TIMING AND ACCURACY TRADE-OFFS

Öztürk, Metin Güney

Master's Program in Electrical and Electronics Engineering

Advisor: Asst. Prof. Dr. Ayça Kumluca Topallı

January, 2026

Early-season crop yield prediction is essential for agricultural planning but remains uncertain due to incomplete observations and evolving crop conditions. This thesis examines the trade-off between forecast timing and yield prediction accuracy within a satellite-based yield forecasting framework, focusing on how uncertainty changes when predictions are issued earlier in the growing season. Multi-year Sentinel-2 imagery was integrated with meteorological data from the Copernicus Climate Data Store to construct weekly time series of vegetation indices and environmental variables. Cloud contamination was addressed through a cloud masking and temporal reconstruction procedure, and the proportion of reconstructed pixels was quantified to assess data reliability. Yield prediction models were developed for forecast months ranging from March to October. Multiple modeling approaches, including Random Forests, Gradient Boosting, and Neural Networks, were evaluated, with Linear Regression selected as the final model due to its robustness, interpretability, and stable

performance across prediction horizons. Model performance was assessed using R^2 , RMSE, MAE, and relative error metrics. Results indicate a strong dependence of prediction accuracy on forecast timing. Early-season predictions exhibit high uncertainty, while performance improves substantially later in the season. July and August provide a practical balance between timeliness and reliability, whereas September and October yield high-confidence estimates. All forecasts target the final end-of-season (December) yield of the corresponding production year.

Keywords: Crop yield prediction, early-season forecasting, satellite remote sensing, Sentinel-2 vegetation indices.



ÖZET

SENTINEL-2 ZAMAN SERİLERİ KULLANILARAK MAKİNE ÖĞRENMESİ TABANLI ERKEN DÖNEM REKOLTE TAHMİNİ: TAHMİN ZAMANI VE DOĞRULUK ARASINDAKİ ÖDÜNLEŞİMİN ANALİZİ

Öztürk, Metin Güney

Elektrik-Elektronik Mühendisliği Yüksek Lisans Programı

Tez Danışmanı: Asst. Prof. Dr. Ayça Kumluca Topallı

Ocak, 2026

Erken dönem rekolte tahmini, tarımsal planlama açısından kritik öneme sahiptir; ancak sınırlı gözlemler ve ürün gelişiminin zaman içinde değişmesi nedeniyle belirsizlik içermektedir. Bu tez, uydu tabanlı bir rekolte tahmin çerçevesi kapsamında, tahmin zamanlaması ile tahmin doğruluğu arasındaki ödünleşimi incelemekte ve büyüme sezonunun erken safhalarında yapılan tahminlerde belirsizliğin nasıl değiştiğine odaklanmaktadır. Çok yıllık Sentinel-2 uydu görüntüleri, Copernicus Climate Data Store'dan elde edilen meteorolojik verilerle birleştirilerek bitki örtüsü indisleri ve çevresel değişkenlerden oluşan haftalık zaman serileri oluşturulmuştur. Bulut kaynaklı bozulmalar, bulut maskeleme ve zamansal yeniden yapılandırma (rekonstrüksiyon) prosedürü ile giderilmiş; yeniden yapılandırılan piksel oranı, veri güvenilirliğini değerlendirmek amacıyla nicel olarak analiz edilmiştir. Rekolte tahmin modelleri, Mart'tan Ekim'e kadar uzanan farklı tahmin ayları için geliştirilmiştir. Rastgele Ormanlar, Gradyan Artırma ve Yapay Sinir Ağları dâhil olmak üzere çeşitli modelleme yaklaşımları değerlendirilmiş; sağlamlığı, yorumlanabilirliği ve farklı

tahmin ufukları boyunca istikrarlı performansı nedeniyle nihai model olarak Doğrusal Regresyon tercih edilmiştir. Model performansı, R^2 , RMSE, MAE ve bağıl hata metrikleri kullanılarak değerlendirilmiştir. Elde edilen bulgular, tahmin doğruluğunun tahmin zamanlamasına güçlü biçimde bağlı olduğunu göstermektedir. Erken dönem tahminleri yüksek belirsizlik sergilerken, sezon ilerledikçe tahmin performansı belirgin şekilde artmaktadır. Temmuz ve Ağustos ayları zamanlılık ile güvenilirlik arasında pratik bir denge sunarken, Eylül ve Ekim ayları yüksek güven düzeyine sahip rekolte tahminleri sağlamaktadır. Tüm tahminler, ilgili üretim yılına ait sezon sonu (Aralık) nihai rekoltesini hedeflemektedir.

Anahtar Kelimeler: Rekolte tahmini, erken dönem tahmin, uydu uzaktan algılama, Sentinel-2 vejetasyon indisleri.

TABLE OF CONTENTS

ABSTRACT.....	iv
ÖZET.....	vi
LIST OF TABLES	xi
LIST OF FIGURES	xii
CHAPTER 1: INTRODUCTION	1
CHAPTER 2: LITERATURE SURVEY.....	4
2.1. Machine Learning Approaches for Yield Prediction	4
2.2. Remote Sensing and Field-Scale Prediction	5
2.3. Climate Drivers, Data Constraints, and Temporal Predictability.....	7
2.4. Positioning the Present Study Within the Literature.....	8
CHAPTER 3: TECHNICAL BACKGROUND	10
3.1. Remote Sensing for Agricultural Monitoring.....	10
3.1.1 Satellite-Based Vegetation Monitoring.....	10
3.1.2 Cloud Masking and Temporal Gap-Filling of Satellite Imagery.....	12
3.2. Vegetation Indices	12
3.2.1 Normalized Difference Vegetation Index (NDVI).....	13
3.2.1.1 Mathematical Definition	14
3.2.1.2 Physiological Interpretation	15
3.2.1.3 Use in Yield Prediction.....	15
3.2.2 Green NDVI (GNDVI)	15
3.2.2.1 Mathematical Definition	16
3.2.2.2 Physiological Interpretation	17
3.2.3 Chlorophyll Red-Edge Index (CHL-RE).....	17
3.2.3.1 Mathematical Definition	18
3.2.3.2 Physiological Interpretation	19
3.2.3.3 Use in Agriculture	19
3.2.4 Normalized Difference Moisture Index (NDMI).....	19
3.2.4.1 Mathematical Definition	21
3.2.4.2 Physiological Interpretation	21
3.2.4.3 Use in Yield Prediction.....	21
3.3. Machine Learning Models.....	22
3.3.1 Linear Regression	22

3.3.1.1 Mathematical Formulation	23
3.3.1.2 Limitations	23
3.3.2 Regularized Linear Models.....	24
3.3.2.1 Ridge Regression.....	24
3.3.2.2 Lasso Regression.....	24
3.3.2.3 Elastic Net.....	24
3.3.2.4 Polynomial Regression.....	24
3.3.3 Random Forest.....	25
3.3.3.1 How It Works	25
3.3.3.2 Advantages for Agriculture	25
3.3.4 Gradient Boosting.....	25
3.3.4.1 Training Mechanism	26
3.3.5 Artificial Neural Networks (NN).....	26
3.3.5.1 Structure.....	27
3.3.5.2 Advantages	27
3.4 Correlation Analysis.....	28
3.5. Model Evaluation Metrics	29
3.5.1 Coefficient of Determination (R^2)	29
3.5.2 RMSE	30
3.5.3 MAE	30
3.5.4 Leave-One-Out Cross-Validation (LOOCV)	30
3.6. Summary Table of Vegetation Indices.....	31
3.7. Summary Table of Machine Learning Models	31
CHAPTER 4: METHODOLOGY	33
4.1 Study Area and Target Crop	34
4.2 Data Sources	34
4.2.1 Sentinel-2 Vegetation Indices	34
4.2.2 Weather Data.....	35
4.3 Preprocessing of Satellite Imagery	36
4.3.1 Cloud Masking Algorithm.....	36
4.3.2 Temporal Interpolation of Masked Pixels	37
4.4 Feature Extraction	38
4.4.1 CHL-RE Features	38
4.4.2 NDVI Features.....	39

4.4.3 GNDVI and NDMI Features.....	39
4.4.4 Weather Features.....	40
4.5 Construction of Seasonal Feature Sets.....	40
4.6 Correlation Analysis and Feature Selection	41
4.7 Machine-Learning Models	42
4.7.1 Linear and Regularized Models.....	42
4.7.2 Tree-Based Model.....	43
4.7.3 Neural Network Models	43
4.8 Hyperparameter Optimization	43
4.9 Cross-Validation and Evaluation Strategy.....	44
4.10 Model Selection	45
4.11 Early-Season Yield Forecasting Framework	45
4.12 Implementation Details	46
CHAPTER 5: RESULTS AND DISCUSSION.....	48
5.1 Impact of Temporal Filling on Data Integrity.....	49
5.2 Model Performance Overview and Implications for Model Selection	52
5.3 Gradient Boosting Results and Comparative Assessment.....	53
5.4 Final Model Selection and Rationale	54
5.5 Prediction Timing vs Accuracy	56
5.5.1 Overall trend: Accuracy improves toward harvest.....	59
5.5.2 March predictions: Very high uncertainty and unstable year-level errors.....	59
5.5.3 April predictions: Moderate improvement, but still low explanatory power	60
5.5.4 May predictions: Not strictly better than April indicative of phenology and noise effects.....	61
5.5.5 June predictions: Presence of a major outlier year and instability despite mid-season timing.....	62
5.5.6 July predictions: Major step up in predictive skill	63
5.5.7 August predictions: Best RMSE/MAE among mid-late months, but R^2 slightly lower than July/September	64
5.5.8 September predictions: Peak stability across months before harvest	65
5.5.9 October predictions: Best overall performance and “reference point” for the thesis trade-off	66
CHAPTER 6: CONCLUSION AND FUTURE WORK	68
REFERENCES.....	71

LIST OF TABLES

Table 1. Vegetation Indices and Sentinel-2 Bands 31

Table 2. ML Algorithms and Characteristics 31

Table 3. Metrics Table of Linear Models..... 54



LIST OF FIGURES

Figure 1. Spectral Bands of the Sentinel-2 Multispectral Instrument (MSI)	11
Figure 2. NDVI Image of Bergama.....	14
Figure 3. Green Normalized Difference Vegetation Index (GNDVI) on 17 December 2024.....	16
Figure 4. Chlorophyll Red-Edge Index (CHL-RE) on 17 December 2024.	18
Figure 5. Normalized Difference Moisture Index (NDMI) on 17 December 2024. ..	20
Figure 6. Visual Representation of Linear Regression	23
Figure 7. Visual Representation of Neural Network Models.....	27
Figure 8. Research Pipeline.....	33
Figure 9. GNDVI Images before cloud masking (a) before filling masked pixels (b) processed image (c).....	38
Figure 10. Future-Past Filling Percentages of Satellite Images	50
Figure 11. Actual vs Predicted Yield on March.....	60
Figure 12. Actual vs Predicted Yield on April.....	61
Figure 13. Actual vs Predicted Yield on May	62
Figure 14. Actual vs Predicted Yield on June.....	63
Figure 15. Actual vs Predicted Yield on July	64
Figure 16. Actual vs Predicted Yield on Augst.....	65
Figure 17. Actual vs Predicted Yield on September	65
Figure 18. Actual vs Predicted Yield on October	66
Figure 19. R2 Score For Each Month	67

CHAPTER 1: INTRODUCTION

Accurate and timely estimation of agricultural yield is a critical component of modern food production systems. Reliable yield forecasts support decision-making across multiple levels, including farm management, supply chain planning, market regulation, and food security policy. Traditionally, yield estimates have relied on field surveys and post-harvest reporting, which, while accurate, are costly, time-consuming, and inherently retrospective. As a result, there has been growing interest in developing predictive methods that can estimate crop yield before harvest, enabling proactive and informed interventions.

With the increasing availability of Earth observation data, remote sensing–based yield prediction has emerged as a powerful alternative to conventional approaches. Satellite platforms such as Sentinel-2 provide frequent, multispectral observations that enable the extraction of vegetation indices related to canopy vigor, chlorophyll content, biomass accumulation, and physiological stress. When combined with meteorological variables such as temperature, precipitation, and radiation, these data offer a rich representation of crop development over time. Advances in machine learning and statistical modeling have further expanded the potential of these datasets, allowing complex relationships between environmental conditions and yield outcomes to be learned from historical records.

Despite these advances, a fundamental challenge remains unresolved: yield prediction accuracy is strongly dependent on prediction timing. Early-season forecasts must operate under incomplete information, as critical phenological stages have not yet occurred and observational data are limited. Consequently, predictions generated early in the growing season often exhibit high uncertainty, while predictions made closer to harvest are more accurate but less useful for early decision-making. This creates an inherent trade-off between timeliness and accuracy, which is central to the practical application of yield forecasting systems.

While many previous studies focus on maximizing prediction accuracy at a single forecast point often near harvest relatively fewer works systematically examine how

prediction error evolves as forecasts are issued progressively earlier in the season. Moreover, early-season yield prediction is further complicated by data quality issues in optical satellite imagery, particularly cloud contamination, which introduces missing or unreliable observations. To address this, cloud masking and gap-filling techniques are commonly applied, but these procedures inevitably alter the original data and may influence prediction performance, especially when large portions of early-season imagery are reconstructed.

Within this context, the primary objective of this thesis is to quantify the trade-off between prediction timing and yield estimation accuracy in a satellite-driven yield forecasting framework. Accordingly, this thesis addresses the following research question: How does the accuracy of crop yield predictions vary as forecasts are generated at progressively earlier stages of the growing season, and at what stage does the reduction in accuracy limit the practical usefulness of early-season predictions? More specifically, the study examines how prediction uncertainty evolves in relation to seasonal data availability and evaluates the extent to which preprocessing procedures, including cloud masking and temporal filling, influence the reliability of satellite-based yield forecasts. Rather than optimizing for maximum accuracy alone, this study aims to understand *when* yield predictions become reliable and how uncertainty decreases as more seasonal information becomes available. By explicitly evaluating prediction performance at multiple forecast months throughout the growing season, this work provides a temporal perspective on yield predictability that is directly relevant to operational agricultural decision-making.

To achieve this objective, multi-year Sentinel-2 satellite imagery was integrated with meteorological data obtained from the Copernicus Climate Data Store. Vegetation indices were derived on a weekly basis and processed using a cloud masking and temporal filling strategy to ensure continuous time series suitable for modeling. The extent of temporal filling was explicitly quantified to assess its potential impact on data integrity, particularly during early-season periods when cloud-related gaps are more prevalent. Yield prediction models were then trained and evaluated at multiple forecast dates ranging from early spring to harvest time.

A range of modeling approaches was explored, including ensemble methods and neural networks, in order to assess their suitability for this task. Through systematic

evaluation, Linear Regression was ultimately selected as the final modeling framework due to its robustness, interpretability, and stable performance across prediction horizons. This choice reflects the methodological emphasis of the thesis: isolating the effects of data availability and seasonal progression on prediction accuracy, rather than introducing confounding variability from overly complex models.

The contributions of this thesis can be summarized as follows. First, it provides a structured and quantitative analysis of how yield prediction accuracy evolves across the growing season using satellite-derived inputs. Second, it evaluates the role of cloud masking and temporal filling in shaping early-season data reliability. Third, it demonstrates that simpler, transparent models can be effective for analyzing timing-based prediction trade-offs when data availability is limited. Finally, it offers practical insights into the earliest points in the season at which yield predictions become sufficiently reliable for operational use.

The remainder of this thesis is organized as follows. Chapter 2 reviews related work in remote sensing-based yield prediction and early-season forecasting. Chapter 3 describes the data sources, preprocessing steps, and cloud masking and filling methodology. Chapter 4 presents the modeling framework and performance evaluation metrics. Chapter 5 analyzes prediction results across different forecast months and discusses the observed trade-offs between timing and accuracy. Finally, Chapter 6 concludes the thesis by summarizing key findings, outlining limitations, and suggesting directions for future research, including the integration of UAV imagery and local sensor networks for enhanced yield forecasting.

CHAPTER 2: LITERATURE SURVEY

Accurate prediction of crop yield is central to modern precision agriculture, enabling farmers, policymakers, and industry stakeholders to optimize resource use, manage risks, and strategically plan agricultural activities. Recent advances in remote sensing, meteorological data availability, and machine learning (ML) techniques have significantly expanded the methodological toolbox for yield forecasting. While deep learning (DL) methods have gained prominence in image-rich environments, much of the contemporary literature demonstrates that classical ML methods remain highly effective often outperforming DL when yield datasets are small, when remote-sensing inputs are aggregated, or when explainability is essential. This survey synthesizes findings from a core set of recent studies and highlights how the present research aligns with and extends current knowledge on data-efficient, ML-based crop yield prediction.

2.1. Machine Learning Approaches for Yield Prediction

Several studies provide broad overviews of ML techniques applied to crop yield forecasting. Van Klompenburg et al. (2020) conducted one of the first systematic literature reviews focusing exclusively on ML for crop yield prediction, covering studies from various countries and crops. Their findings reveal that Random Forest (RF), Support Vector Regression (SVR), Gradient Boosting Machines (GBM), and simple linear regression remain widely used due to their robustness, generalizability, and compatibility with heterogeneous datasets that include weather, soil, and vegetation indices. Reported R^2 values across reviewed studies frequently fall in the 0.6–0.9 range, particularly when multi-source environmental data are included.

In a complementary review, Chlingaryan et al. (2018) emphasized precision-agriculture-oriented ML applications, notably nitrogen and crop-status assessment from hyperspectral reflectance, multispectral remote sensing, and proximal sensing. Their survey demonstrates that ML more than DL dominates scenarios where data availability or spatial resolution is limited, highlighting the importance of engineered features and agronomic domain knowledge in achieving reliable performance.

Beyond reviews, multiple comparative studies illustrate the practical strengths of ML models. Kim et al. (2019) evaluated ANN, RF, SVR, and other AI models for predicting yields of major crops in the Midwestern United States (2006–2015). Their analysis shows that ML models often outperform traditional regression and can rival shallow neural networks when climate and management variables are aggregated at seasonal or sub-seasonal resolution.

At broader scales, Paudel et al. (2021) applied GBM to sub-national European yield forecasting using the MCYFS (MARS Crop Yield Forecasting System) database. Their model, trained on historical weather and yield data from numerous EU countries, produced normalized RMSE (nRMSE) values competitive with operational monitoring systems. Importantly, they demonstrated that even early-season predictions retain useful accuracy an observation directly relevant to studies, like the present one, that explore the limits of seasonal predictability.

Meanwhile, meta-modeling approaches such as those used by Xu et al. (2021) illustrate how ML can approximate process-based simulations. In their study across China, RF and SVR approximated the outputs of the WheatGrow crop model using minimal environmental data such as latitude, DEM, and a few temperature-based indices. This reduction in data requirements is highly relevant for data-constrained environments. And achieved RMSE scores ranged between 1400-1800 kg/ha for different ML models.

Finally, Zhao et al. (2015) explored how the spatial and temporal granularity of meteorological data affects regional-scale crop modeling. Their analysis concluded that both overly coarse and overly fine weather data can degrade yield prediction performance, underscoring the need for appropriate aggregation a principle echoed by the methodologies of many ML studies.

2.2. Remote Sensing and Field-Scale Prediction

Remote sensing data, particularly vegetation indices like NDVI, GNDVI, SAVI, EVI, and red-edge metrics, provide critical information about crop phenology and health. However, the literature shows a clear division between deep-learning image-based approaches and engineering-based, aggregated-index ML approaches.

Filippi et al. (2019) developed an ML system for grain yield prediction across multiple farms in Western Australia using multi-layered datasets that included combine harvester yield maps, terrain characteristics, soil conductivity, and climate indicators. Using RF-based ensembles, their models achieved RMSE values of roughly 0.36–0.42 t ha⁻¹ and concordance correlation coefficients up to ~0.9 later in the season, demonstrating the strong predictive value of engineered remote-sensing and environmental features.

Arshad et al. (2023) conducted one of the most detailed ML studies integrating satellite-derived vegetation indices and climate variables for wheat yield prediction in Pakistan. Using scenarios constructed from NDVI, GNDVI, EVI, SAVI, and comprehensive weather variables, they found that RF achieved the best performance, with post-season $R^2 = 0.783$, $RMSE \approx 2.07$ q ha⁻¹, and $MAE \approx 1.68$ q ha⁻¹. The study highlights that even when advanced indices are available, aggregated spectral representations rather than pixel-level CNN features combined with weather data provide strong performance in real-world agricultural systems.

Similarly, Leo et al. (2021) predicted within-field cotton yields using public remote-sensing (Landsat-derived GNDVI), climate data (rainfall), and soil properties across commercial farms in New South Wales, Australia. Their models, trained using 10-fold and temporal leave-one-year-out cross-validation, achieved R^2 up to 0.52 and $RMSE \sim 0.15$ – 0.20 t ha⁻¹. Their finding that prediction accuracy significantly declines for unseen years echoes a key principle explored in the conducted research: yield models must be evaluated in temporal generalization settings rather than random cross-validation.

Abbas et al. (2020) applied ML algorithms (RF, SVR) to proximal sensing reflectance data of sorghum, maize, and millet in smallholder systems. Their models achieved RMSE as low as 0.19–0.29 t ha⁻¹ and $RPD > 2.5$, demonstrating how vegetation reflectance whether proximal or satellite-derived can serve as a strong predictor of yield when aggregated appropriately.

Gaussian Process regression has also been applied to yield prediction. Abbas et al. (2021) predicted potato yields in the Netherlands using GP models trained on gridded weather and soil data. Reported RMSE values ranged 4.62–6.93 t ha⁻¹, and GP models

offered interpretable temporal response curves an important consideration for operational forecasting.

The ecological distance algorithm developed by Li Tian et al. (2020) predicted rice and wheat yields in China by leveraging climate and soil similarity across regions. While non-ML, this approach aligns with ML philosophies of environmental analogs and multi-source covariate integration.

ML-driven multimodal studies also extend beyond remote sensing. Torsoni et al. (2022) estimated soybean yield across Brazil using climate-driven ML models, confirming that seasonal climate features alone can capture a substantial portion of interannual variability.

2.3. Climate Drivers, Data Constraints, and Temporal Predictability

Large-scale analyses, such as Hoffman et al. (2018) on sub-Saharan Africa, reveal how long-term climate signals and warming trends affect realized yields. Their RF models show modest but measurable climate sensitivities ($\approx 10 \text{ kg ha}^{-1}$ per $^{\circ}\text{C}$ increased temperature), while illustrating that management and technological improvements explain more variance than climate alone.

These findings resonate strongly with studies examining environmental data resolution requirements. Zhao et al. showed that meteorological data aggregation can significantly affect performance, while Xu et al. demonstrated that a minimal set of environmental factors can achieve high accuracy in ML meta-modeling.

Collectively, these works highlight two key realities for ML-based crop yield prediction:

- (1) Data scarcity is common, especially for field-level yield measurements;
- (2) Prediction performance depends strongly on seasonal timing, with uncertainty increasing as predictions are made earlier in the crop cycle.

2.4. Positioning the Present Study Within the Literature

The present thesis aligns closely with the dominant trends identified in the literature, yet addresses a less-explored but practically critical question: How far into the season can accurate yield predictions be made when only aggregated remote-sensing indices and weather data are available?

Several characteristics of the research's methodology position it firmly within best practices:

(i) Aggregated remote sensing inputs.

Much like Filippi et al., Arshad et al., Leo et al., and Abbas et al., our approach uses median and average NDVI derived spectral values rather than pixel-level CNN based extraction. This strategy is widely validated in the literature for small datasets or heterogeneous fields, reducing noise and computational requirements.

(ii) Integration of weather and vegetation indices.

The work's feature set follows the strongest performing multimodal approaches, as demonstrated by Arshad (2023), Paudel (2021), and Xu (2021), all of whom show that combining weather with spectral summaries boosts predictive performance.

(iii) Evaluation under data-scarce conditions.

Deep learning methods frequently underperform when yield data are limited as noted by multiple reviews and case studies. The observation is that DNNs performed poorly, while simple models like linear regression achieved the best results, matches the consensus across these works.

(iv) Temporal generalization and leave-one-out validation.

By applying leave-one-out evaluation across years and testing prediction accuracy at multiple forecast horizons (e.g., July, August, September, ...), the research adopts the most rigorous validation framework identified in the literature similar to the temporal leave-one-year-out strategy of Leo et al. (2021) and the early/late-season comparison of Paudel et al.

(v) Contribution to understanding predictability limits.

While many studies maximize predictive accuracy, few systematically analyze the limits of predictability given realistic data constraints. The research contributes

uniquely to the literature by explicitly quantifying how accuracy deteriorates as predictions are made earlier in the season.



CHAPTER 3: TECHNICAL BACKGROUND

Modern agricultural yield prediction relies on the integration of remote sensing, vegetation indices, and machine learning models capable of learning non-linear relationships from high-dimensional temporal data. With the availability of high-resolution satellite platforms such as Sentinel-2, it is possible to extract spectral information that captures canopy vigor, chlorophyll content, biomass accumulation, and physiological stress. These rich spectral datasets enable the calculation of vegetation indices such as the Normalized Difference Vegetation Index (NDVI), Green NDVI (GNDVI), and Chlorophyll Red-Edge Index (CHL-RE) metrics that have become essential components of predictive agronomic modeling.

In parallel, machine learning (ML) and statistical modeling techniques such as RFs, NN, Artificial NNs, and penalized linear regression models (Ridge, Lasso, ElasticNet) have evolved to process these multidimensional feature sets. These models enable the extraction of latent structures embedded within temporal sequences of satellite-derived and meteorological data. Evaluation metrics including Coefficient of Determination (R^2), Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) play a central role in quantifying predictive accuracy and model generalization.

This chapter provides the technical background required to understand the methodological foundations of the models and indices used in this thesis. Each subsection integrates mathematical definitions, historical context, applications in agricultural research, and citations to widely recognized studies in the field.

3.1. Remote Sensing for Agricultural Monitoring

3.1.1 Satellite-Based Vegetation Monitoring

Remote sensing has revolutionized agronomic research by providing continuous, objective, and large-scale measurements of crop development. The European Space Agency's Sentinel-2 mission provides multispectral imagery at 10–20 m spatial

resolution and 4-day revisit frequency, making it one of the most valuable resources for agricultural yield modeling.

Sentinel-2 carries the Multispectral Instrument (MSI), which measures surface reflectance across 13 spectral bands. Of particular relevance for vegetative monitoring are the:

- Red band (Band 4, 665 nm; 10 m)
- Near-infrared (NIR) band (Band 8, 842 nm; 10 m)
- Green band (Band 3, 560 nm; 10 m)
- Red-edge bands (Bands 5, 6, 7, 705–783 nm; 20 m)

According to the European Space Agency (ESA, 2015) and Drusch et al. (2012), the spectral bands shown in Fig. 1 facilitate the calculation of vegetation indices sensitive to chlorophyll content, structural canopy properties, photosynthetic activity, and biomass accumulation.

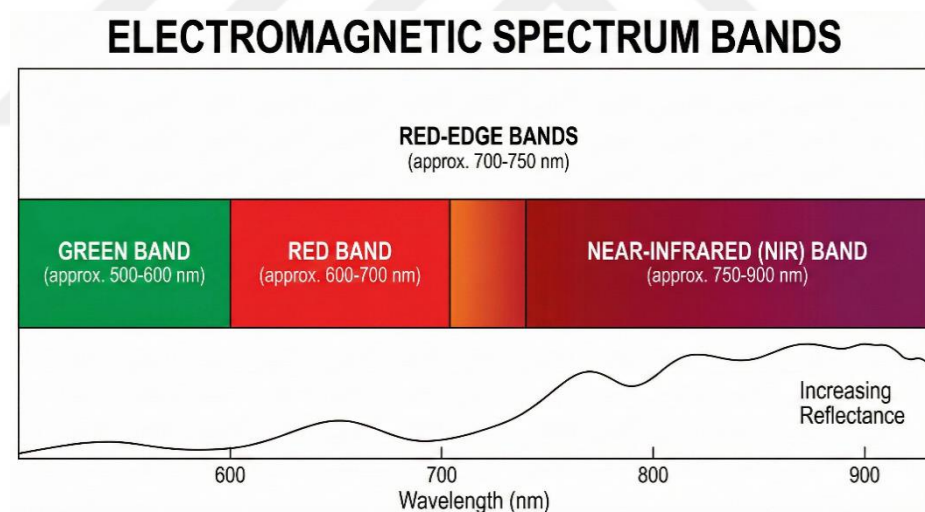


Figure 1. Spectral Bands of the Sentinel-2 Multispectral Instrument (MSI)

Vegetation indices are dimensionless transformations of spectral reflectance values designed to enhance vegetation signals while minimizing confounding factors such as soil background, illumination differences, and atmospheric interference.

3.1.2 Cloud Masking and Temporal Gap-Filling of Satellite Imagery

Optical satellite imagery is frequently contaminated by clouds, haze, and cloud shadows, which obscure the land surface and break the continuity of vegetation index time series. For yield prediction, these gaps are problematic because machine learning models rely on temporally consistent spectral signals to capture crop development and stress patterns. To address this, a cloud-masking and temporal gap-filling step is typically performed before extracting vegetation indices and model features.

In practice, cloud masking is implemented by using per-pixel quality flags and spectral thresholds to identify cloudy or otherwise contaminated pixels, which are then excluded from further analysis. Han et al. [16] applied a cloud mask to all Landsat 7 and Sentinel-2 images over sugarcane blocks in Queensland, Australia, before calculating vegetation and landscape indices such as NDVI, EVI, Chlorophyll Index Green (CI), NDMI, Bare Soil Index (BI) and NDBI. This ensured that the derived indices represented actual canopy and surface conditions rather than atmospheric artefacts.

After cloud masking, residual temporal gaps are often filled using compositing or simple temporal interpolation strategies. Following the approach of Han et al. [16], seasonal composites can be constructed (e.g., a “previous harvest” composite and a “current season” composite), where each composite uses the latest available cloud-free image before a specified cut-off date. In cases where the target date is cloudy, missing pixels are replaced with values from the nearest preceding clear observation. This procedure yields spatially complete, seasonally representative covariates suitable for early- and late-season yield forecasting. In this thesis, a similar strategy is adopted to generate gap-reduced Sentinel-2 NDVI, GNDVI and CHL-RE products, providing temporally coherent inputs to the machine learning models.

3.2. Vegetation Indices

Vegetation indices serve as proxies for plant physiological traits. They reduce multidimensional spectral information into scalar metrics that reflect vegetation vigor, chlorophyll absorption, and canopy structure.

The vegetation indices used in this work include:

- NDVI – Normalized Difference Vegetation Index
- GNDVI – Green Normalized Difference Vegetation Index
- CHL-Red-Edge (RECI or Chlorophyll Index – Red Edge)
- NDMI (Normalized Difference Moisture Index)

Each section explains the mathematical formulation, spectral rationale, and their relevance for crop yield prediction.

3.2.1 Normalized Difference Vegetation Index (NDVI)

NDVI reflects overall canopy vigor by comparing visible-red absorption to NIR reflectance. High NDVI areas indicate green, dense, photosynthetically active vegetation, whereas lower values represent sparse cover, senescent vegetation, bare soil or built-up surfaces. In Fig. 2. this late-season observation, spatial differences in NDVI highlight variability in biomass accumulation and canopy health across the study area, supporting its use as a predictor of yield potential.

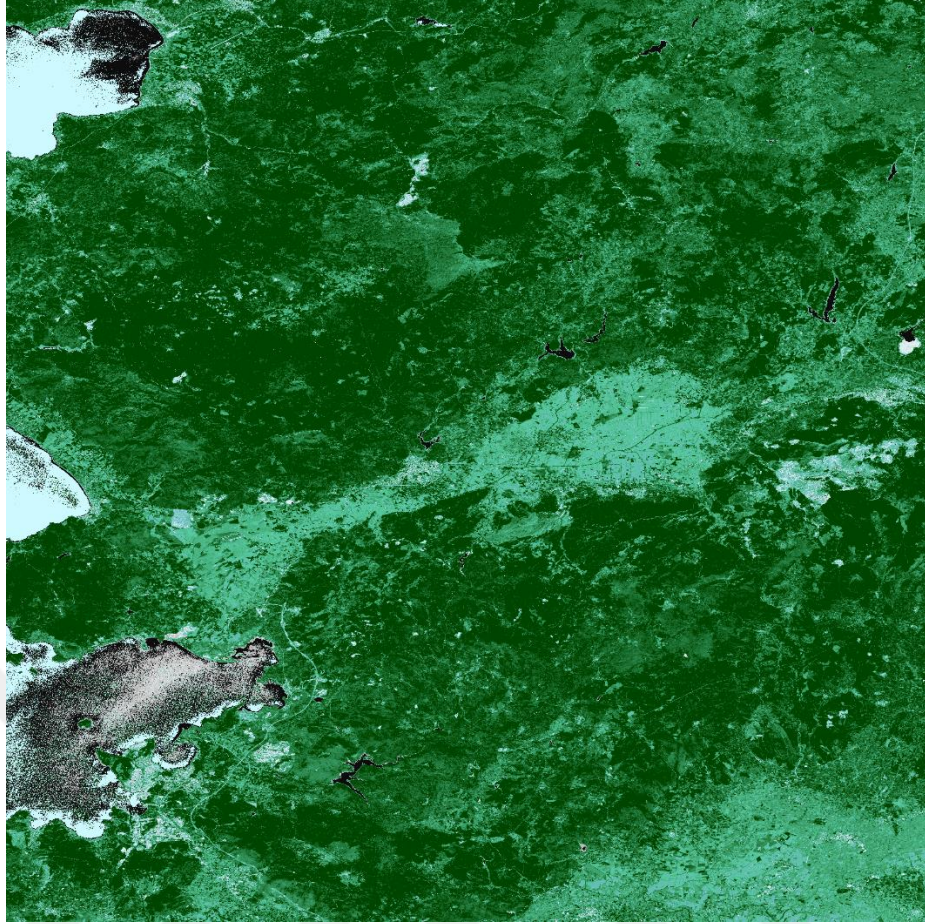


Figure 2. NDVI Image of Bergama

NDVI is the most widely used vegetation index in agricultural and ecological monitoring. First introduced by Rouse et al. (1974), NDVI measures the contrast between red and near-infrared reflectance of vegetation.

3.2.1.1 Mathematical Definition

$$NDVI = \frac{NIR - RED}{NIR + RED} \quad (1)$$

Where:

- NIR = Sentinel-2 Band 8 (842 nm)
- RED = Sentinel-2 Band 4 (665 nm)

3.2.1.2 Physiological Interpretation

Healthy vegetation absorbs most red light (for photosynthesis) and reflects a large portion of NIR due to cellular structure. NDVI values range from:

- -1 to 0 → water, clouds, built-up areas
- 0.2–0.5 → sparse vegetation
- 0.6–0.9 → dense, healthy vegetation

3.2.1.3 Use in Yield Prediction

NDVI correlates with:

- Leaf Area Index (LAI)
- Chlorophyll content
- Biomass
- Canopy density
- Soil moisture deficiency

Numerous studies have demonstrated that NDVI time series can explain a substantial proportion of yield variability for crops such as maize, wheat, soybean, and rice (Tucker, 1979; Wardlow & Egbert, 2008).

3.2.2 Green NDVI (GNDVI)

In Fig. 3, the GNDVI layer shows strong spatial variability in crop chlorophyll content at the end of the season. High GNDVI areas (bright pixels) indicate dense canopy with high leaf nitrogen concentration, while darker areas represent lower chlorophyll levels linked to nutrient stress, reduced canopy biomass, or late-stage senescence. Because GNDVI is more sensitive to nitrogen status than NDVI, the map highlights zones where nutrient management or yield potential differs within the study area.

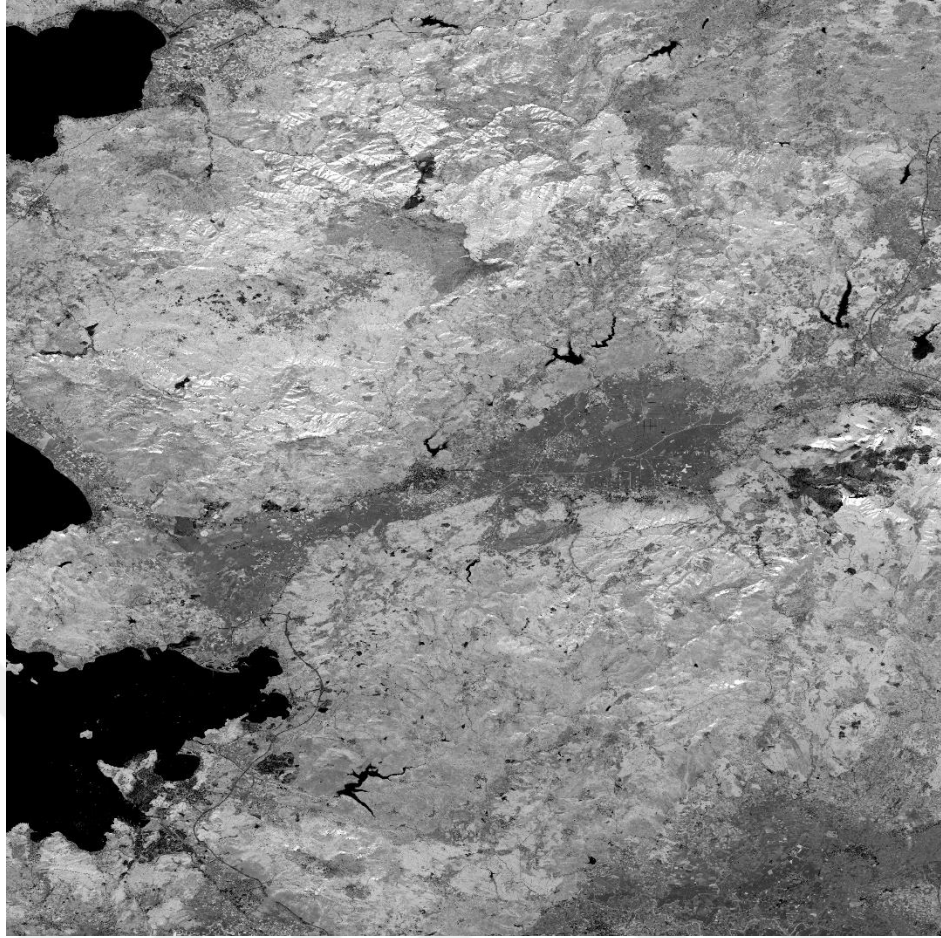


Figure 3. Green Normalized Difference Vegetation Index (GNDVI) on 17 December 2024.

GNDVI, introduced by Gitelson and Merzlyak (1996), replaces the red band with the green band, making it more sensitive to chlorophyll concentration.

3.2.2.1 Mathematical Definition

$$GNDVI = \frac{NIR - GREEN}{NIR + GREEN} \quad (2)$$

Where:

- NIR = Sentinel-2 Band 8
- GREEN = Sentinel-2 Band 3 (560 nm)

3.2.2.2 Physiological Interpretation

GNDVI is more responsive than NDVI in cases where:

- Crops have high LAI (NDVI saturation occurs)
- Nutrient-related stress affects chlorophyll content
- Early- to mid-season chlorophyll accumulation is critical

Studies show that GNDVI is a stronger predictor of nitrogen status, photosynthetic rates, and early phenological development (Gitelson, 2004; Haboudane et al., 2004).

3.2.3 Chlorophyll Red-Edge Index (CHL-RE)

This red-edge-based chlorophyll index highlights biochemical variations in the canopy. As seen in Fig. 4, high CHL-RE values indicate areas with high pigment concentration and vigorous photosynthetic activity, whereas darker tones capture zones of reduced chlorophyll, early stress, nutrient deficiencies, or structural thinning of the canopy. Compared with GNDVI, CHL-RE provides greater sensitivity to subtle changes at the end of the season, making it useful for detecting early stress that precedes yield reduction.

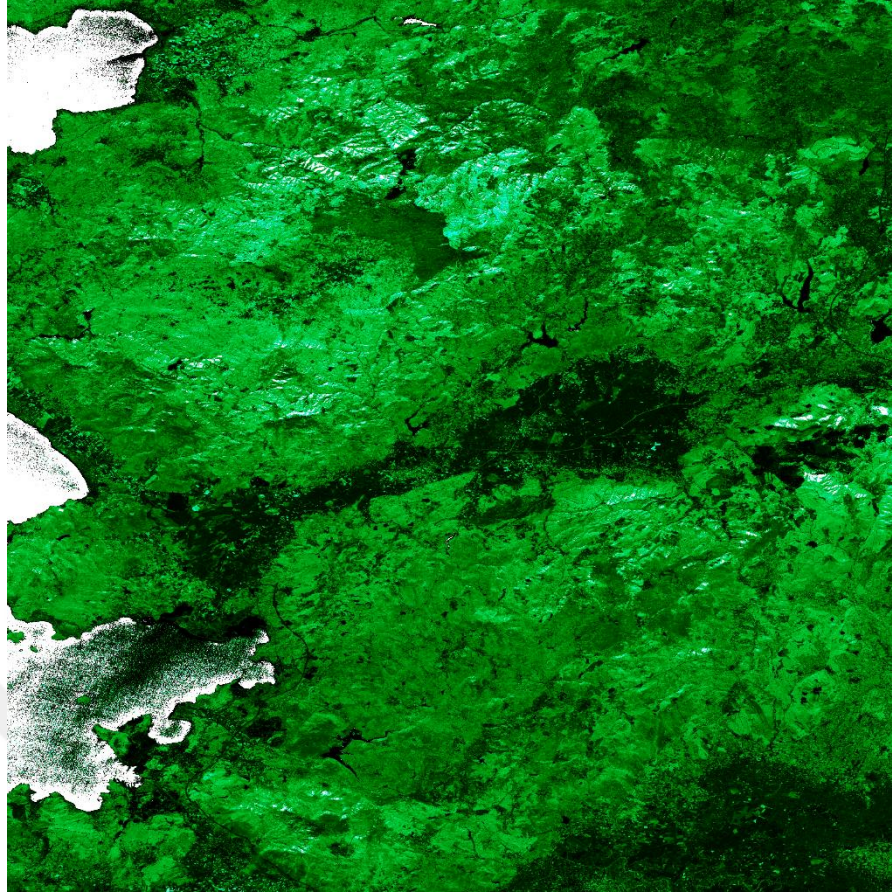


Figure 4. Chlorophyll Red-Edge Index (CHL-RE) on 17 December 2024.

The chlorophyll red-edge index (also called the Red-Edge Chlorophyll Index, RECI) exploits Sentinel-2's unique red-edge bands (705–783 nm), which are highly sensitive to leaf chlorophyll concentration.

3.2.3.1 Mathematical Definition

A commonly used form (Gitelson et al., 2003):

$$\text{CHL_RE} = \frac{\text{NIR}}{\text{RE}} - 1 \quad (3)$$

Where:

- NIR = Band 8 (842 nm)
- RE = Band 5 (705 nm) or Band 6/7 depending on the chosen variant

3.2.3.2 Physiological Interpretation

The red-edge region corresponds to the transition between strong chlorophyll absorption (red) and high NIR reflectance. It is extremely sensitive to:

- Chlorophyll concentration
- Leaf biochemical traits
- Canopy structural changes
- Nutrient deficiencies

3.2.3.3 Use in Agriculture

CHL-RE is widely used for:

- Early stress detection
- Nitrogen mapping
- Predicting yield in cereals, maize, soybean, and tobacco
- Discriminating vegetation from soil background

Its sensitivity makes it particularly powerful for yield modeling in crops with high pigment variation (Dash & Curran, 2004; Gitelson et al., 2003).

3.2.4 Normalized Difference Moisture Index (NDMI)

Fig. 5. illustrates the spatial distribution of the Normalized Difference Moisture Index (NDMI) over the study area on 17 December 2024.

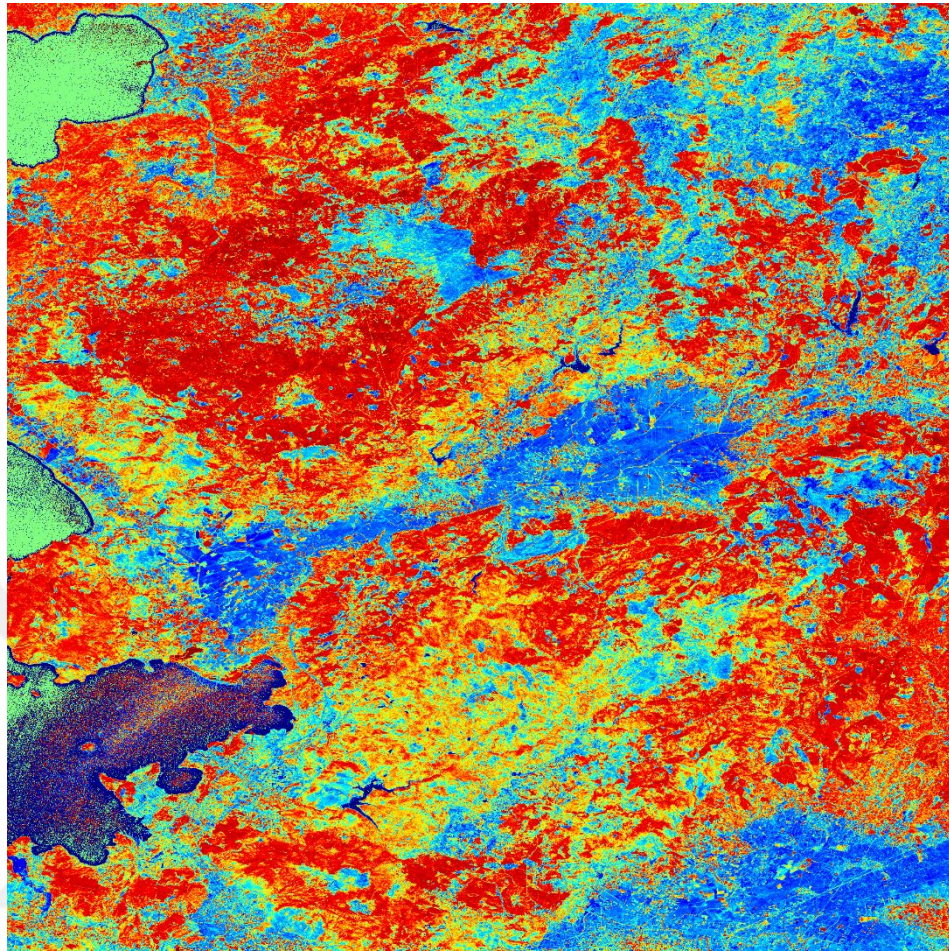


Figure 5. Normalized Difference Moisture Index (NDMI) on 17 December 2024.

NDMI estimates canopy moisture content using NIR and SWIR reflectance. High NDMI values indicate well-hydrated vegetation with high leaf water content, while lower values represent moisture stress, senescence, or exposed soil. The spatial variability in NDMI across the study area reflects differences in crop water status and soil moisture retention, which are critical factors influencing biomass production and yield.

The Normalized Difference Moisture Index (NDMI) is a spectral indicator used to assess vegetation water content and canopy hydration levels. Introduced by Gao (1996), NDMI exploits the difference in reflectance between the near-infrared (NIR) and shortwave-infrared (SWIR) spectral regions, where SWIR reflectance is strongly affected by liquid water within leaves.

3.2.4.1 Mathematical Definition

$$\text{NDMI} = \frac{\text{NIR} - \text{SWIR}}{\text{NIR} + \text{SWIR}} \quad (4)$$

Where:

- NIR = Sentinel-2 Band 8 (842 nm)
- SWIR = Sentinel-2 Band 11 (1610 nm) or Band 12 (2190 nm) depending on application.

3.2.4.2 Physiological Interpretation

NDMI increases with:

- higher vegetation water content
- dense and healthy canopy biomass
- reduced moisture stress

Low NDMI values indicate drought stress, low canopy hydration, or sparse vegetation cover. Because water stress limits biomass accumulation and photosynthetic activity, NDMI serves as a useful proxy for yield-limiting environmental stress (Gao, 1996; Pandey et al., 2019).

3.2.4.3 Use in Yield Prediction

NDMI has been successfully applied in agricultural monitoring to:

- detect early moisture stress
- improve crop growth modeling
- enhance the predictive power of vegetation-index-based yield models

Studies show that combining NDMI with chlorophyll-sensitive indices such as NDVI, GNDVI and CHL-RE improves the prediction of yield variability caused by drought and irrigation differences [18], [19]. In this thesis, NDMI is utilized alongside other indices to better represent water-driven yield variability within the study area.

3.3. Machine Learning Models

Modern crop yield prediction frameworks employ a range of machine learning algorithms capable of modeling complex, nonlinear interactions between vegetation indices, weather variables, soil parameters, and temporal features.

This thesis utilizes:

- Random Forest
- Gradient Boosting
- Neural Networks
- Linear Regression
- Ridge
- Lasso
- ElasticNet
- Polynomial Regression

Each method is explained in depth.

3.3.1 Linear Regression

Linear regression based models were employed to capture relationships between input variables and the target response under the assumption of linear dependence. These models offer a transparent and analytically tractable framework in which the contribution of each predictor can be directly interpreted through its estimated coefficient. Regularization techniques were incorporated to control model complexity and improve numerical stability when predictors exhibit correlation or differing magnitudes. Figure 6 presents a conceptual illustration of the linear modeling framework, highlighting the role of coefficient estimation and regularization in the prediction process.

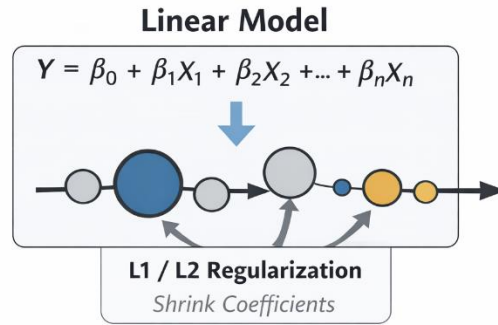


Figure 6. Visual Representation of Linear Regression

3.3.1.1 Mathematical Formulation

Linear regression models the relationship between dependent variable y and predictors $x_1, x_2, x_3, \dots, x_n$:

$$y = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_n x_n + \varepsilon \quad (5)$$

Where:

- β = coefficients
- ε = residual error

3.3.1.2 Limitations

Linear regression assumes:

- Linearity
- Normality of errors
- No multicollinearity
- Homoscedasticity

In remote sensing-based yield modeling, linear regression remains valuable because it provides a clear, interpretable baseline that helps quantify the strength and direction of relationships between vegetation indices and yield.

3.3.2 Regularized Linear Models

Regularization improves generalization by penalizing large coefficients.

3.3.2.1 Ridge Regression

Ridge introduces an L2 penalty:

$$\min_{\beta} (|y - X\beta|^2 + \lambda|\beta|_2^2) \quad (6)$$

Reduces variance, handles multicollinearity, but does not enforce sparsity.

3.3.2.2 Lasso Regression

Lasso introduces an L1 penalty:

$$\min_{\beta} (|y - X\beta|^2 + \lambda|\beta|_1) \quad (7)$$

Advantages:

- Feature selection (coefficients shrink to zero)
- Interpretable models

3.3.2.3 Elastic Net

Combines L1 and L2 penalties:

$$\min_{\beta} (|y - X\beta|^2 + \alpha\lambda|\beta|_1 + (1 - \alpha)\lambda|\beta|_2^2) \quad (8)$$

Useful when predictors are correlated common in vegetation index datasets.

3.3.2.4 Polynomial Regression

To model nonlinearity:

$$y = \beta_0 + \beta_1x + \beta_2x^2 + \dots + \beta_dx^d \quad (9)$$

Polynomial regression captures curvature but risks overfitting if not regularized.

3.3.3 Random Forest

RF, introduced by Breiman (2001), is an ensemble learning algorithm based on multiple decision trees.

3.3.3.1 How It Works

RF constructs:

- Multiple decision trees from bootstrapped samples
- Random subsets of features at each split

Prediction is given by:

$$\hat{y} = \frac{1}{N} \sum_{i=1}^N T_i(x) \quad (10)$$

Where T_i is the i -th tree.

3.3.3.2 Advantages for Agriculture

- Handles nonlinear relationships
- Robust to noise
- Works well with multicollinearity
- Provides feature importance
- Resistant to overfitting

RF is one of the most widely used models in yield prediction literature (Jeong et al., 2016; Johnson, 2016).

3.3.4 Gradient Boosting

GBM, introduced by Friedman (2001), sequentially build trees that correct the errors of previous ones.

3.3.4.1 Training Mechanism

Each new tree fits the gradient of the loss function:

$$y_i^{(m)} = y_i^{(m-1)} - \eta \cdot g_i \quad (11)$$

Where:

- g_i = gradient of loss
- η = learning rate

Popular variants include:

- XGBoost
- LightGBM
- CatBoost

GBM is extremely powerful for remote sensing regression tasks due to its ability to model complex interactions.

3.3.5 Artificial Neural Networks (NN)

Neural network based models were employed to represent nonlinear relationships between input features and the target variable through layered compositions of affine transformations and nonlinear activation functions. These models extend linear formulations by enabling hierarchical feature interactions to be learned directly from the data. Both shallow and deep architectures were considered, differing primarily in network depth and representational capacity. Figure 7 provides a schematic illustration of the neural network modeling framework, highlighting the structure of hidden layers, activation functions, and the flow of information from input to output.

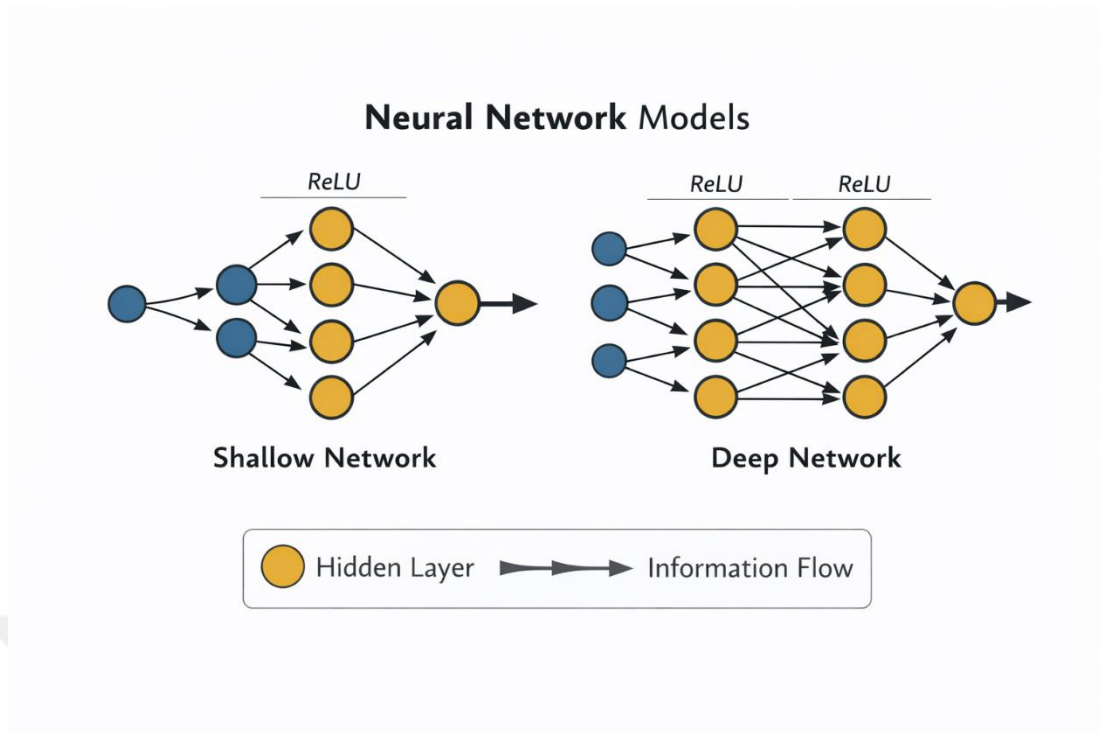


Figure 7. Visual Representation of Neural Network Models

Neural Networks are universal function approximators capable of modeling highly nonlinear patterns.

3.3.5.1 Structure

A typical multilayer perceptron:

$$h^{(1)} = f(W_1x + b_1) \quad (12)$$

$$h^{(2)} = f(W_2h^{(1)} + b_2) \quad (13)$$

$$\hat{y} = W_3h^{(2)} + b_3 \quad (14)$$

3.3.5.2 Advantages

- Capture nonlinearity
- Learn interactions without explicit feature engineering
- Flexible architectures

Neural networks have been applied extensively in yield estimation, especially with large-scale remote sensing datasets (Khaki & Wang, 2019).

3.4 Correlation Analysis

Correlation analysis is a fundamental exploratory technique used to quantify the strength and direction of linear relationships between predictor variables and crop yield. In remote sensing-based agronomic modeling, vegetation indices and meteorological variables often exhibit strong multicollinearity due to their shared physiological and environmental drivers. For instance, NDVI, GNDVI, and CHL-Red-Edge (RECI) values all respond to chlorophyll content, photosynthetic capacity, and canopy biomass, making correlation analysis essential for diagnosing redundant features before model training.

A standard measure is the Pearson correlation coefficient (r), defined as:

$$r = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum(x_i - \bar{x})^2 \sum(y_i - \bar{y})^2}} \quad (15)$$

Values close to ± 1 indicate strong linear associations, while values near 0 represent weak or no linear relationship. In agricultural yield prediction, correlation analysis serves two major purposes:

1. Feature Selection and Dimensionality Reduction

By identifying highly correlated predictor groups, correlation matrices help prevent model instability and inflated variance caused by multicollinearity. Regularized models such as Ridge, Lasso, and ElasticNet directly benefit from reduced redundancy, improving generalization performance. Studies have shown that vegetation indices often cluster into correlated groups reflecting shared biophysical properties, making feature filtering a necessary preprocessing step in yield modeling pipelines (Liu et al., 2020; Potgieter et al., 2016).

2. Identification of Dominant Predictors

Correlation trends across the growing season allow researchers to detect phenological windows where satellite-derived indices or weather variables most strongly influence yield outcomes. For example, mid-season NDVI and red-edge chlorophyll indices typically show peak correlation with final grain yield due to their sensitivity to biomass accumulation and nitrogen status (Peng

et al., 2018; Bolton & Friedl, 2013). This temporal insight informs both model design and agronomic interpretation.

Correlation analysis is routinely used in crop yield prediction studies to:

- remove redundant spectral indices,
- identify key phenological stages,
- guide model complexity, and
- improve interpretability of machine learning pipelines.

Its integration into this thesis ensures that the feature set used for predictive modeling captures the most informative vegetation and environmental signals while mitigating risks associated with multicollinearity.

3.5. Model Evaluation Metrics

Evaluating model accuracy is essential to compare predictive performance across years and regions.

The research uses:

- Coefficient of Determination (R^2)
- Root Mean Squared Error (RMSE)
- Mean Absolute Error (MAE)
- Leave-One-Out Cross-Validation (LOOCV)

3.5.1 Coefficient of Determination (R^2)

$$R^2 = 1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2} \quad (16)$$

Represents the proportion of variance explained.

Interpretation

- 1.0 → perfect prediction
- 0 → no better than mean

- Can be negative → worse than baseline

3.5.2 RMSE

$$\text{RMSE} = \sqrt{\frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2} \quad (17)$$

Sensitive to large errors important when yield values vary widely across seasons.

3.5.3 MAE

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (18)$$

More robust to outliers.

3.5.4 Leave-One-Out Cross-Validation (LOOCV)

Leave-One-Out Cross-Validation (LOOCV) is a robust evaluation strategy commonly used in crop yield forecasting to assess the temporal generalization ability of predictive models. Instead of randomly splitting the dataset, LOOCV sequentially removes a single observation (or a full season in temporal applications) and trains the model on all remaining samples. The model then predicts the left-out case, and the process is repeated until every observation has been excluded once.

This approach is particularly valuable in agricultural modeling because it mimics real forecasting conditions in which a future unseen season must be predicted using only historical data. By preventing temporal leakage between training and test data, Leave-One-Out Cross-Validation (LOOCV) provides a more realistic estimate of model performance and reduces overly optimistic accuracy scores often observed in random train–test splits (Hastie et al., 2009).

In yield prediction studies, LOOCV has been used to rigorously validate machine learning models that rely on vegetation indices, weather variables, and block-level agronomic data (Han et al., 2022), making it especially suitable for the temporally structured datasets employed in this thesis.

3.6. Summary Table of Vegetation Indices

Table 1. Vegetation Indices and Sentinel-2 Bands

Index	Formula	Sentinel-2 Bands	Sensitivity
NDVI	$(\text{NIR} - \text{Red}) / (\text{NIR} + \text{Red})$	B8, B4	Biomass, canopy vigor
GNDVI	$(\text{NIR} - \text{Green}) / (\text{NIR} + \text{Green})$	B8, B3	Chlorophyll, nitrogen
CHL-RE (RECI)	$(\text{NIR} / \text{RE}) - 1$	B8, B5/6/7	Chlorophyll, early stress

3.7. Summary Table of Machine Learning Models

Table 2. ML Algorithms and Characteristics

Model	Type	Strengths	Limitations
Linear Regression	Parametric	Simple, interpretable	Assumes linearity
Ridge	Parametric	Handles collinearity	No sparsity
Lasso	Parametric	Feature selection	Struggles with high collinearity
ElasticNet	Parametric	Balanced	Requires tuning α
Polynomial	Parametric	Models curvature	Overfitting risk
Random Forest	Ensemble	Robust, nonlinear	Less interpretable
Gradient Boosting	Ensemble	High accuracy	Sensitive to tuning

Table 2 (continued)

Neural Network	Deep learning	Flexible, powerful	Requires large data
----------------	---------------	--------------------	---------------------

This technical background chapter provided a comprehensive overview of the vegetation indices, machine learning algorithms, and evaluation metrics used in this thesis. NDVI, GNDVI, and CHL-Red-Edge derived from Sentinel-2 MSI play a crucial role in quantifying crop vigor and biochemical status. Machine learning models including RF, GB, and NNs offer powerful frameworks for extracting latent nonlinear relationships between spectral features and yield outcomes. Regularized linear regression models improve interpretability and robustness when dealing with correlated vegetation indices.

The combination of remote sensing and machine learning constitutes a strong foundation for modern agricultural yield estimation and forms the backbone of the methodological approach used in this work.

CHAPTER 4: METHODOLOGY

This chapter describes the methodological framework developed to forecast tobacco yield in the Bergama region using multi-temporal Sentinel-2 vegetation indices, reanalysis-based weather variables, and machine-learning models. The workflow follows the pipeline illustrated in Figure 8 (data acquisition → cloud masking → interpolation of masked pixels → feature extraction → correlation-based feature selection → model training and hyperparameter optimization → early-season yield prediction). The methodological steps implemented in this chapter are grounded in the theoretical and technical foundations summarized in Chapter 3. In particular, Sentinel-2 data characteristics and index derivations are described in 3.1-3.2, preprocessing concepts (cloud masking and gap filling) are introduced 3.1.2, and the modeling and evaluation framework is detailed in 3.3-3.4.



Figure 8. Research Pipeline

4.1 Study Area and Target Crop

The study area is the Bergama district in western Türkiye, an agriculturally important region located in İzmir province. The analysis was conducted over a rectangular extent defined in WGS84 coordinates by the bounding box:

$$(26.8108^{\circ} \text{ E}, 39.5951^{\circ} \text{ N}, 27.6457^{\circ} \text{ E}, 38.6469^{\circ} \text{ N}) \quad (19)$$

The target crop is tobacco, which is widely cultivated in the region and is characterized by relatively stable management practices and limited rotational cropping. Tobacco's tendency to be grown repeatedly on similar land parcels reduces the impact of crop rotation on inter-annual yield variability, making it suitable for temporal yield modeling at regional scale.

Annual tobacco yield statistics for Bergama for the years 2017–2024 were obtained from the Turkish Statistical Institute (TÜİK). Yields are reported at district level in tons per hectare (t ha^{-1}) and constitute the reference values for model training and evaluation.

4.2 Data Sources

4.2.1 Sentinel-2 Vegetation Indices

Multi-temporal satellite data were acquired from the Sentinel-2 Level-2A (S2 L2A) product, which provides atmospherically corrected surface reflectance. Data access and processing were performed using the SentinelHub Python interface, with a bounding box matching the Bergama study area. To harmonize processing and reduce computational load, a target spatial resolution of 1 m was defined and then capped to a maximum raster size of 2500×2500 pixels, ensuring that all years are processed at a consistent spatial extent.

From Sentinel-2, four vegetation indices were computed:

- Normalized Difference Vegetation Index (NDVI)
- Green NDVI (GNDVI)
- Chlorophyll Red-Edge Index (CHL-RE)

- Normalized Difference Moisture Index (NDMI)

The spectral rationale and mathematical definitions of NDVI, GNDVI, CHL-RE, and NDMI are provided in the Technical Background (3.2.1-3.2.4), while Sentinel-2 MSI band characteristics relevant to these indices are summarized in 3.1.1.

Each index was produced as a weekly composite, resulting in approximately four images per month and 52 images per year for each index. Weekly dates were aligned across years to maintain a comparable temporal structure.

Custom evalscripts were used to compute the indices on the fly. For example, the CHL-RE index was defined using Sentinel-2 red-edge and NIR bands (B05, B07) as:

$$\text{CHL} - \text{RE} = \frac{B07}{B05} - 1 \quad (20)$$

with the resulting index visualized using a colorBlend mapping in the evalscript. Similar scripts were used to compute NDVI, NDMI, and GNDVI, following standard formulations in the remote sensing literature.

4.2.2 Weather Data

Meteorological variables representing near-surface temperature, humidity, precipitation, solar radiation and wind conditions were obtained from Copernicus reanalysis products (e.g., AgERA5). The temporal resolution of the original data varies (hourly or daily depending on variable), but all variables were aggregated to a daily or sub-daily representation and then aligned with the weekly time steps used for satellite imagery.

The following weather features were considered:

- 2 m dew point temperature
- Cloud cover mean
- Derived relative humidity (2 m) maximum and minimum (24 h)
- Precipitation duration fraction
- Precipitation flux
- Precipitation rain duration fraction

- Precipitation solid duration fraction
- Relative humidity at 2 m for 06:00, 12:00, and 18:00
- Solar radiation flux
- 2 m air temperature: maximum (24 h), maximum during day, mean (24 h), mean during day, mean during night, minimum (24 h), minimum during night
- Mean 10 m wind speed

All weather variables were spatially averaged over the Bergama bounding box so that each weekly time step is represented by a set of area-mean meteorological indicators.

4.3 Preprocessing of Satellite Imagery

The need for preprocessing in optical imagery particularly cloud contamination and temporal discontinuities is discussed in 3.1.2. The following steps operationalize those concepts for the Bergama Sentinel-2 time series.

4.3.1 Cloud Masking Algorithm

Cloud contamination is a major challenge in optical remote sensing. To mitigate this, a cloud masking algorithm was applied to each Sentinel-2 scene before feature extraction. Cloud and cloud-shadow pixels were identified using a threshold-based approach; pixels whose index or reflectance-based composite exceeded a pre-defined threshold value ($th = 50$ in the scaled representation used in the pipeline) were marked as cloudy. This procedure implements the general cloud-masking principles introduced in the Technical Background (3.1.2), translating them into a threshold-based masking rule consistent with the scaled representation used in this pipeline.

The algorithm operates on each weekly index image of NDVI, NDMI, GNDVI, and CHL-RE. Pixels flagged as cloudy are set to “no data” in the corresponding index raster. This step ensures that subsequent feature extraction is based only on valid vegetation signals and not dominated by spurious values from clouds or atmospheric artifacts.

The conceptual structure of this step is consistent with cloud-masking procedures adopted in recent yield-forecasting work, where reliable masking is a prerequisite for temporal interpolation and feature derivation.

4.3.2 Temporal Interpolation of Masked Pixels

Cloud masking introduces spatial gaps in the time series. To reconstruct continuous spatiotemporal signals, a temporal interpolation procedure was implemented. The motivation and common strategies for temporal gap filling in cloud-affected optical time series are outlined in 3.1.2; here, they are implemented as a nearest-neighbor temporal fill applied at the pixel level across the weekly stacks. For each pixel location within a given year and index:

1. The algorithm scans the 52 weekly images for that year.
2. If a pixel is missing (cloud-masked) at a given week, the algorithm searches the nearest available week in time, either in the preceding or following weeks.
3. The missing value is filled using that nearest neighbor in the temporal dimension.
4. The procedure iterates until all missing pixels for that year are filled.

As a result, each year's stack of weekly NDVI, GNDVI, CHL-RE, and NDMI images is fully populated, with no missing pixels. This allows robust derivation of temporal statistics and avoids biases caused by uneven cloudiness across years.

Figure 9 Illustration of the satellite image preprocessing workflow applied in this study using a weekly GNDVI composite over the Bergama region: (a) original Sentinel-2 GNDVI image affected by cloud contamination, (b) cloud-masked image after applying the threshold-based cloud masking algorithm ($th = 50$), and (c) temporally interpolated image obtained by filling cloud-masked pixels using the nearest valid observation in the temporal domain.

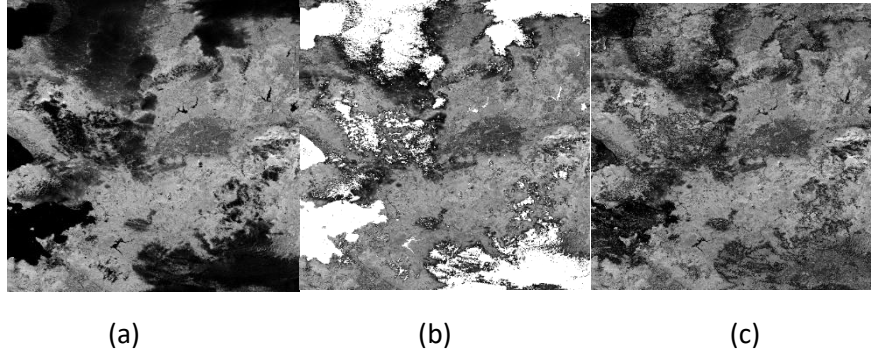


Figure 9. GNDVI Images before cloud masking (a) before filling masked pixels (b) processed image (c)

4.4 Feature Extraction

From the cloud-free, temporally interpolated index stacks and the aligned weather series, a set of aggregated image-level and weather-level features was computed for each weekly time step. The aim is to summarize the spatial distribution of vegetation and moisture status over Bergama in a compact feature vector suitable for machine learning. This feature design is informed by the index interpretations in 3.2 (physiological meaning of NDVI/GNDVI/CHL-RE/NDMI) and the modeling requirements described in 3.3 (learning from multivariate predictors under limited samples).

4.4.1 CHL-RE Features

For CHL-RE images (stored as RGBA composites), the following features were extracted:

- red_mean, red_std
- green_mean, green_std
- blue_mean, blue_std
- alpha_mean, alpha_std
- brightness_mean, brightness_std (overall luminance/intensity)
- corr_rg, corr_rb, corr_gb (pairwise color-channel correlations)
- red_visible_mean, green_visible_mean, blue_visible_mean (means computed only over visible pixels)

These features capture both spectral intensity and color relationships, which are related to chlorophyll content, canopy vigor, and spatial heterogeneity.

4.4.2 NDVI Features

For NDVI composites (also represented in RGB form after color mapping), an analogous feature set was derived:

- red_mean, red_std
- green_mean, green_std
- blue_mean, blue_std
- alpha_mean, alpha_std
- brightness_mean, brightness_std
- corr_rg, corr_rb, corr_gb
- visible_pixel_count
- red_visible_mean, green_visible_mean, blue_visible_mean

These statistics summarize the overall distribution of NDVI intensities and their spatial patterns.

4.4.3 GNDVI and NDMI Features

GNDVI and NDMI were stored as grayscale images. For each weekly image, the following features were computed:

- gray_mean
- gray_std
- gray_min
- gray_max
- pixel_count (number of valid pixels)

GNDVI features capture the canopy's green reflectance relative to NIR, which is often more sensitive to chlorophyll in certain crops, while NDMI features reflect the canopy moisture status.

4.4.4 Weather Features

For each weekly time step, the weather dataset was represented by the following variables:

- 2mdewpointtemperature_data
- Cloud_Cover_Mean
- Derived_Relative_Humidity_2m_Max_24h
- Derived_Relative_Humidity_2m_Min_24h
- Precipitation_Duration_Fraction
- Precipitation_Flux
- Precipitation_Rain_Duration_Fraction
- Precipitation_Solid_Duration_Fraction
- Relative_Humidity_2m_06h
- Relative_Humidity_2m_12h
- Relative_Humidity_2m_18h
- Solar_Radiation_Flux
- Temperature_Air_2m_Max_24h
- Temperature_Air_2m_Max_Day_Time
- Temperature_Air_2m_Mean_24h
- Temperature_Air_2m_Mean_Day_Time
- Temperature_Air_2m_Mean_Night_Time
- Temperature_Air_2m_Min_24h
- Temperature_Air_2m_Min_Night_Time
- Wind_Speed_10m_Mean

Together with the vegetation index features, these variables form a multisource environmental feature set that describes both the state of the canopy and the atmospheric conditions affecting growth.

4.5 Construction of Seasonal Feature Sets

For each year, weekly features were first computed for all indices and weather variables. To build the predictors used for yield modeling, cumulative seasonal feature sets were assembled month by month.

For a given forecast month m (e.g., March, April, ..., November) in year y :

- All features from the start of the growing season up to the last week within month m were combined into a fixed-length vector.
- Only information available up to month mm was used; no data from later months entered the feature set.

Thus, for each year and month, a feature vector summarizing the partial-season trajectory of vegetation and weather conditions was produced, with the corresponding annual tobacco yield in t ha^{-1} as the target value.

This procedure results in a series of forecasting tasks:

- “March prediction”: using information up to March.
- “April prediction”: using information up to April.
- ...
- “November prediction”: nearly full-season information.

The number of training samples for each task equals the number of years (2017–2024), and feature dimensionality increases with month as more weeks are included.

4.6 Correlation Analysis and Feature Selection

Given the potentially large number of candidate predictors (all vegetation-index statistics and weather variables at multiple time steps), a correlation-based feature selection step was applied to reduce dimensionality and focus on the most informative features. The role of correlation analysis in diagnosing multicollinearity and guiding feature reduction for remote sensing predictors is discussed in the Technical Background (3.3.6). The present study applies that rationale by selecting the top 20 predictors by absolute Pearson correlation within each forecast month.

Let X_{final} denote the full feature matrix for a given forecast month, with rows corresponding to years and columns to candidate features, and let y_{final} be the vector of observed yields. For each feature j , the absolute Pearson correlation with yield was computed:

$$corr_j = |corr(X \cdot j, y_{final})| \quad (21)$$

The 20 features with the highest absolute correlation with yield were retained as the input for model training. This approach balances interpretability and parsimony, ensuring that the models focus on the most yield-sensitive indicators while avoiding excessive dimensionality given the limited number of years.

4.7 Machine-Learning Models

To explore different functional forms between environmental predictors and yield, a diverse set of regression models was trained for each forecast month using the selected top 20 features. The theoretical motivation and core formulations of the model families used here are summarized in 3.3 (linear and regularized regression, polynomial regression, RF, GB, and neural networks). This section focuses on their implementation and tuning within the forecasting framework.

4.7.1 Linear and Regularized Models

1. Ordinary Least Squares Linear Regression
 - a. Provides a baseline assuming a linear relationship between predictors and yield.
2. Ridge Regression
 - a. Adds L2 regularization, penalizing large coefficients and improving stability in the presence of multicollinearity.
3. Lasso Regression
 - a. Uses L1 regularization, encouraging sparse solutions and performing implicit feature selection.
4. Elastic Net
 - a. Combines L1 and L2 penalties, balancing sparsity and coefficient shrinkage.
5. Polynomial Regression (Poly)
 - a. Extends the linear model by including polynomial transformations of the input features (e.g., degree-2 interactions), enabling simple nonlinear relationships to be captured.

4.7.2 Tree-Based Model

6. Random Forest Regressor

- a. An ensemble of decision trees trained with bootstrap aggregation and random feature selection.
- b. Capable of modeling nonlinear interactions and handling complex feature spaces with relatively low risk of overfitting when tuned appropriately.

4.7.3 Neural Network Models

7. Shallow Neural Networks

- a. Multi-layer perceptrons (MLPs) with one or two hidden layers.
- b. Designed to capture nonlinear relationships while remaining relatively simple given the modest sample size.

8. Deep Neural Networks

- a. MLP architectures with multiple hidden layers and higher capacity, allowing more complex mappings between features and yield.

Neural networks used nonlinear activation functions (e.g., ReLU) and were trained with backpropagation using stochastic gradient descent variants (e.g., Adam optimizer). Regularization (such as dropout or weight decay) and early stopping were employed as needed to mitigate overfitting.

4.8 Hyperparameter Optimization

For each model type, hyperparameter tuning was carried out using a combination of Grid Search and Randomized Search within the scikit-learn framework. This tuning step supports the model capacity generalization trade-off described in 3.3, ensuring that each algorithm's complexity is adapted to the available seasonal information and limited number of annual samples.

- RandomizedSearchCV was first used to explore a broad hyperparameter space efficiently (e.g., number of trees and max depth for RF, regularization strengths for linear models, architecture and learning rate for neural networks).
- For promising regions of this space, GridSearchCV was subsequently applied to refine hyperparameters on a finer grid.

Hyperparameter search was performed separately for each forecast month and model type, reflecting the fact that the optimal complexity may change as the amount of available information grows through the season.

4.9 Cross-Validation and Evaluation Strategy

The total number of available years (2017–2024) is small, so a cross-validation strategy that makes maximal use of the data is necessary. The rationale for LOOCV in temporally structured yield forecasting, and the interpretation of evaluation metrics (R^2 , RMSE, MAE), are detailed in 3.5. This study adopts LOOCV to emulate the operational setting in which each year must be predicted using the remaining historical seasons.

For each forecast month:

1. For each year y in $\{2017, \dots, 2024\}$:
 - a. The year y is held out as a test sample.
 - b. All other years are used as training data.
 - c. Hyperparameter optimization is performed using only the training set.
 - d. The model with the best cross-validated performance is refit on all training years and used to predict the yield for year y .
2. This process yields one prediction per year for each model and month.
3. The predictive performance of each model is quantified using metrics such as:
 - a. Coefficient of determination R^2
 - b. Root Mean Squared Error (RMSE)
 - c. Optionally Mean Absolute Error (MAE) or Mean Absolute Percentage Error (MAPE)

LOOCV is particularly suitable in this setting because it evaluates models under realistic conditions of temporal generalization: each year is predicted only from information learned from other years, which mirrors the operational scenario in which future yields are predicted from past seasons.

4.10 Model Selection

For each forecast month and each model type, the LOOCV predictions were collected and evaluated. The best performing model for a given month was defined as the model configuration (algorithm + hyperparameters) that achieved the most favorable combination of high R^2 low RMSE over the 2017–2024 period.

Produced best model used for prediction through season:

- March prediction
- April prediction
- ...
- November prediction

This best model summarize how the optimal modeling strategy evolves as more of the growing season becomes observable. For instance, early in the season simple regularized linear models may be adequate, whereas later months may favor nonlinear models (e.g., RF or neural networks) when more temporal information is available.

4.11 Early-Season Yield Forecasting Framework

The methodology is explicitly designed to support early-season yield forecasting. For each month, the model uses only those features that would be known at that point in an operational setting. This month-by-month forecasting definition follows the evaluation perspective introduced in 3.5, where predictive skill is assessed as a function of information availability through the season.

Formally, let $Fm(y)$ denote the feature vector for year y constructed from all satellite and weather data up to month m . Let $f_m(\cdot)$ be the best model selected for month m . The predicted yield for year y at month m is:

$$\hat{Y}_{y,m} = f_m(F_m(y)) \quad (22)$$

with the true yield Y_y obtained from TÜİK. Because no information after month m is used, this setup quantifies how early in the season reliable forecasts can be produced, and how forecast accuracy increases as the season progresses.

4.12 Implementation Details

All data preparation and modeling steps were implemented in Python 3.11 on a Windows workstation. A typical software stack included:

- NumPy and Pandas for data manipulation and feature engineering
- SentinelHub-Py and associated libraries for Sentinel-2 data access and evalscript execution
- Scikit-learn for machine-learning models, hyperparameter tuning (GridSearchCV, RandomizedSearchCV), and evaluation
- TensorFlow/Keras (or an equivalent deep-learning library) for the implementation of shallow and deep neural networks
- Matplotlib for visualization and exploratory analysis

Computations were run on a standard CPU-based environment with sufficient RAM (16–32 GB) to handle the 2500×2500 pixel imagery and generated feature tables.

The methodology integrates:

1. Region-specific yield statistics for Bergama tobacco (2017–2024),
2. Cloud-masked and temporally interpolated Sentinel-2 vegetation indices (NDVI, GNDVI, CHL-RE, NDMI),
3. Aggregated weather indicators from reanalysis products,
4. Rich feature extraction capturing spatial, temporal, and spectral characteristics,
5. Correlation-based feature selection to retain the most yield-sensitive predictors,
6. A diverse portfolio of machine-learning models (linear, regularized, tree-based and neural networks) with systematic hyperparameter optimization,
7. Leave-One-Out Cross-Validation to rigorously assess temporal generalization, and
8. Month-by-month forecasting experiments to evaluate early-season yield predictability.

This framework provides a robust basis for examining how well tobacco yields in Bergama can be forecast from remote sensing and weather information, how predictive performance varies across the season, and how the methodology compares to approaches reported in the existing literature.



CHAPTER 5: RESULTS AND DISCUSSION

The primary objective of this thesis is to quantify the tradeoff between prediction timing and yield estimation accuracy in remote sensing-based crop yield forecasting. In operational agricultural monitoring, yield predictions are often required before the end of the growing season, when complete observational data are not yet available. However, early season predictions inherently rely on incomplete information and therefore introduce uncertainty. Understanding how prediction error evolves as forecasts are generated further away from harvest is critical for evaluating the reliability, robustness, and practical usability of machine learning based yield prediction systems.

To investigate this trade off, this study integrates Sentinel 2 optical satellite imagery and meteorological variables obtained from the Copernicus Climate Data Store, constructing a temporally consistent multi year dataset. Given the well known sensitivity of optical satellite data to cloud contamination, a dedicated cloud masking and temporal filling procedure was applied prior to feature extraction. Cloud affected pixels were identified and replaced using information from either preceding or subsequent valid observations in the time series, enabling the construction of continuous weekly vegetation index datasets. This step is essential to prevent the propagation of missing or spurious values into downstream yield models, particularly in time dependent learning frameworks.

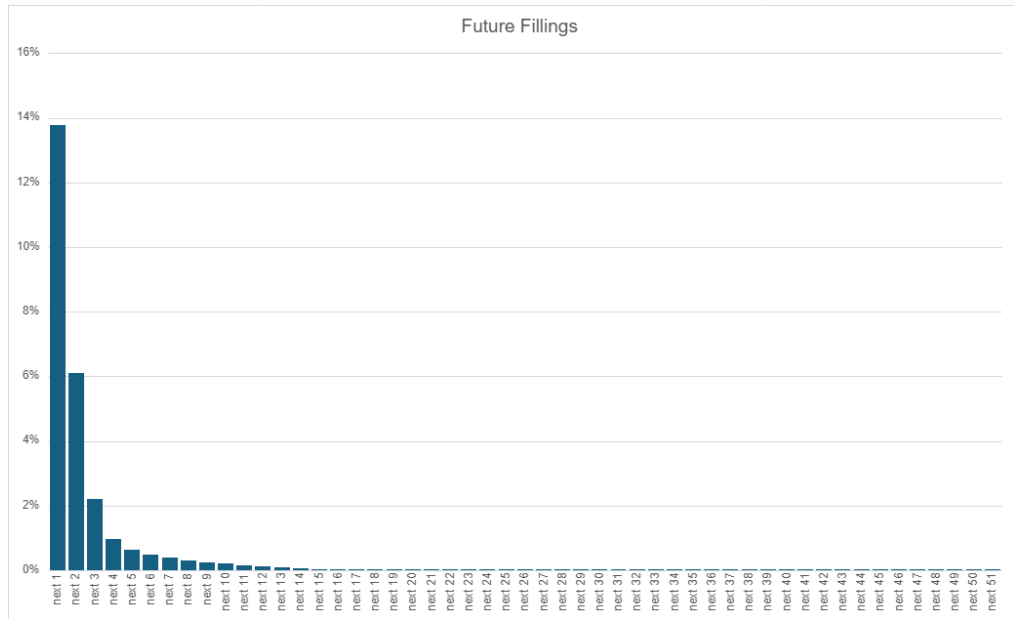
The results presented in this chapter therefore reflect not only the performance of yield prediction models at different forecasting horizons, but also the structural properties of the reconstructed input data. Figures illustrating past based and future based pixel fillings summarize the average proportion of pixels replaced at each temporal offset, providing insight into how much information is synthetically reconstructed as one move away from the observation date. Past fillings represent cases where missing pixels are replaced using information from previous weeks, while future fillings represent replacements derived from subsequent observations. These distributions reveal how heavily the dataset relies on temporal interpolation at different stages of the season.

Importantly, the proportion and direction of pixel replacement have direct implications for yield prediction accuracy. Early season predictions, by definition, depend more strongly on filled data, as fewer valid satellite observations are available. As the prediction date moves closer to harvest, the reliance on filled pixels decreases, and the input features increasingly reflect direct observations rather than reconstructed values. Consequently, any systematic distortion introduced by temporal filling is expected to have a non-uniform impact across prediction horizons.

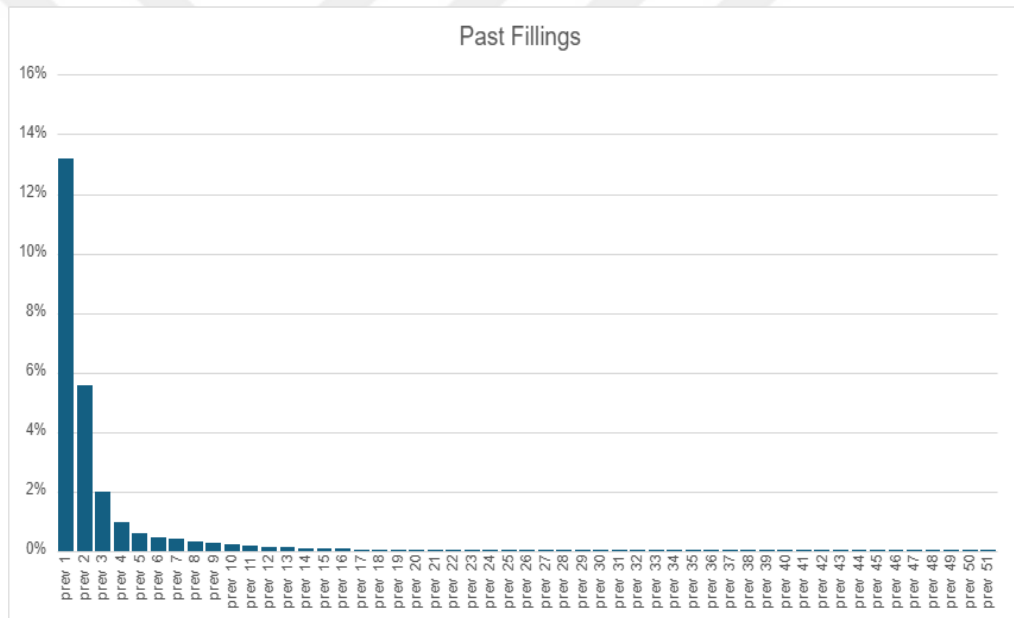
This chapter therefore examines the results from two complementary perspectives. First, it presents the empirical behavior of the cloud masking and filling process, quantifying how much of the spatial information is reconstructed from past or future observations at each time step. Second, it relates these findings to the broader question of prediction reliability, forming the basis for a discussion on whether temporal filling introduces meaningful bias or whether its impact diminishes as more observational data become available. By jointly analyzing prediction timing, data reconstruction intensity, and model performance, this chapter provides a structured framework for interpreting the accuracy timeliness trade off in satellite driven yield forecasting.

5.1 Impact of Temporal Filling on Data Integrity

Cloud masking is an unavoidable preprocessing step in optical remote sensing based agricultural monitoring; however, it introduces spatial gaps that must be addressed to maintain temporal continuity. In this study, masked pixels were reconstructed using a temporal filling strategy based on adjacent observations in the time series. To evaluate the extent to which this procedure alters the original data, the proportion of pixels replaced at each temporal offset was quantified separately for past based fillings and future based fillings, as illustrated in Figures 8.a and 1.b.



(a)



(b)

Figure 10. Future-Past Filling Percentages of Satellite Images

Figure 10.a presents the average percentage of pixels replaced using future observations (future fillings), while Figure 10.b shows the corresponding percentages for past based replacements. In both cases, a strongly right skewed distribution is observed, with the highest proportion of replaced pixels occurring at the smallest temporal offsets. Specifically, the first temporal step exhibits a markedly higher replacement ratio compared to subsequent steps, after which the proportion decreases rapidly and stabilizes at very low levels.

This pattern indicates that the majority of reconstructed pixels originate from observations that are temporally close to the missing data. Such behavior is expected in agricultural time series, where cloud contamination often affects short, contiguous periods rather than extended intervals. From a data integrity perspective, this is a favorable characteristic, as short gap temporal interpolation is less likely to introduce unrealistic spectral behavior than long range extrapolation.

A comparison between past and future fillings further reveals that both strategies exhibit similar decay patterns, suggesting that the temporal filling process does not disproportionately rely on information from one direction in time. However, future based fillings show a slightly higher peak at the earliest offset, indicating that, when available, subsequent observations are more frequently used to reconstruct missing pixels. This asymmetry reflects the practical availability of data during post season analysis, where future observations can be incorporated retrospectively, but it also highlights a potential source of information leakage if not carefully controlled in early season prediction scenarios.

Crucially, the rapid decline in replacement percentages beyond the first few temporal steps suggests that the reconstructed dataset remains largely observation driven for most of the season. As the number of valid satellite acquisitions increases, the dependence on temporally filled pixels becomes negligible. This implies that any distortion introduced by the filling process is concentrated in early season periods, where observational scarcity is highest.

From an integrity standpoint, the results indicate that temporal filling primarily acts as a local smoothing mechanism rather than a dominant source of synthetic information. The limited spatial extent of replaced pixels and their reliance on nearby observations reduce the likelihood of large scale structural bias in vegetation indices. Nevertheless, because early season data inherently contain a higher proportion of reconstructed pixels, predictions made at earlier stages of the season are more susceptible to uncertainties arising from the filling process.

These findings establish an important context for interpreting yield prediction performance in subsequent sections. The observed trade off between prediction timing and accuracy must be understood not only in terms of model capability, but also in

relation to the evolving quality and completeness of the underlying input data. The next section therefore examines how these data integrity characteristics translate into prediction error behavior across different forecast horizons.

5.2 Model Performance Overview and Implications for Model Selection

To complement the analysis of data integrity and temporal filling effects, preliminary yield prediction experiments were conducted using both RF and NN models. These models were selected as representative examples of tree-based ensemble learning and nonlinear function approximation, which are commonly employed in remote sensing-based yield prediction studies.

The RF results, summarized across multiple hyperparameter configurations, exhibit unstable and generally weak predictive performance. While certain configurations achieve relatively low mean absolute error, the corresponding coefficient of determination values are frequently negative or close to zero. Negative R^2 values indicate that, for several configurations, the model performs worse than a naive baseline that predicts the mean yield. This behavior suggests that the RF models struggle to capture the underlying yield-feature relationships in the present dataset, potentially due to high dimensionality, temporal dependency in the input features, or limited sample size relative to model complexity.

Similarly, the NN experiments show consistently high error levels across training and validation phases. Although gradual reductions in training loss and MAE are observed over epochs, validation errors remain large and show limited convergence toward acceptable accuracy levels. The persistent gap between loss reduction and practical predictive performance indicates that the neural models are unable to generalize effectively from the available data. This may reflect sensitivity to noise introduced by temporal filling, insufficient sample size for deep learning architectures, or a mismatch between model capacity and the spatiotemporal structure of the problem.

Importantly, these results do not invalidate the broader research objective of this thesis. Instead, they highlight a critical methodological insight: model choice plays a decisive role in quantifying the trade-off between prediction timing and accuracy. The observed performance limitations suggest that neither the tested RF configurations nor the

implemented NN architectures are optimally suited to the specific characteristics of the dataset and forecasting task.

Consequently, these findings motivate a more targeted search for modeling approaches that can better accommodate temporally evolving inputs, multicollinearity among vegetation indices, and limited sample sizes. Subsequent sections therefore focus on identifying and evaluating alternative models that are more appropriate for early-season yield forecasting, ensuring that the observed prediction-time trade-offs are driven by data availability rather than model inadequacy.

5.3 Gradient Boosting Results and Comparative Assessment

In addition to RF and NN models, GB regression was evaluated as an alternative ensemble-based approach capable of capturing nonlinear relationships while maintaining stronger regularization control. GB is particularly well suited to structured tabular data with correlated predictors, which aligns with the characteristics of the vegetation index and meteorological feature set used in this study.

Across a wide range of hyperparameter configurations, the GB models demonstrate substantially improved and more stable performance compared to both RF and NN approaches. The best-performing configurations achieve an R^2 value of approximately 0.59, accompanied by a mean absolute error (MAE) of around 5.6 units and a root mean squared error (RMSE) of approximately 8.35 units. In addition, the explained variance metric exceeds 0.60, indicating that a considerable portion of yield variability is captured by the model.

Notably, the results show high consistency across different numbers of estimators when shallow tree depths (max depth = 2) and moderate learning rates are used. This stability suggests that the model's performance is not driven by overfitting to specific parameter combinations, but rather by a robust underlying relationship between input features and yield. Increasing the number of estimators beyond a certain threshold yields negligible performance gains, indicating convergence toward a stable solution.

The learning rate emerges as a key control parameter. Higher learning rates (0.10–0.12) produce the strongest overall results, while lower learning rates (0.06) lead to noticeable performance degradation despite increased estimator counts. This behavior

reflects the trade-off between incremental learning and model expressiveness, emphasizing the importance of controlled boosting dynamics for this application.

Despite its improved performance, the GB model does not fully resolve the challenges identified earlier in the chapter. While the achieved accuracy represents a meaningful improvement over baseline models, it remains insufficient for definitive early-season yield prediction without further refinement. Importantly, the observed performance should be interpreted in the context of temporally filled input data, where early-season predictions are inherently influenced by a higher proportion of reconstructed pixels.

These findings suggest that GB provides a more appropriate modeling baseline for investigating the trade-off between prediction timing and accuracy. However, further analysis is required to determine whether the observed improvements are consistent across different prediction horizons and whether the model's sensitivity to temporally reconstructed data varies throughout the season. Consequently, subsequent sections focus on refining the modeling framework and evaluating prediction performance as a function of forecast timing rather than solely optimizing global accuracy metrics.

5.4 Final Model Selection and Rationale

Following an extensive evaluation of multiple modeling paradigms, including ensemble-based methods and neural networks, this study concludes the model selection phase with Linear Regression as the most appropriate baseline model for analyzing the trade-off between prediction timing and accuracy. A comparative summary of linear, regularized, and polynomial regression variants is presented in Figure 2.

Table 3. Metrics Table of Linear Models

Model Type	variant	R2	RMSE	MAE
LinearRegression	LinearRegression_v1	0.7035	7.1010	5.8546
LinearRegression	LinearRegression_v0	0.6789	7.3913	6.1634
Lasso	Lasso_a0.001_v0	0.6855	7.3146	6.0882

Table 3 (continued)

Ridge	Ridge_a1.0_v1	0.6689	7.5055	6.1643
Ridge	Ridge_a0.1_v0	0.6778	7.4038	6.1654
ElasticNet	ElasticNet_a0.01_l10.1_v0	0.6766	7.4172	6.2166
ElasticNet	ElasticNet_a0.1_l10.1_v3	0.6689	7.5049	6.2263
ElasticNet	ElasticNet_a0.01_l10.5_v1	0.6633	7.5685	6.5515
Poly2	Poly2_d2_v0	0.6765	7.4184	6.0404
Poly3	Poly3_d3_v0	0.6696	7.4967	5.9098

As shown in Table 3, the Linear Regression model achieves the highest coefficient of determination ($R^2 = 0.704$) among all tested variants, while maintaining a comparatively low root mean squared error (RMSE = 7.10) and mean absolute error (MAE = 5.85). Regularized alternatives such as Ridge, Lasso, and ElasticNet exhibit slightly reduced explanatory power, indicating that additional penalization does not improve generalization for this dataset. Similarly, higher-order polynomial models fail to outperform the linear baseline, suggesting that increased model complexity does not yield meaningful gains.

These results indicate that the relationship between aggregated vegetation indices, meteorological variables, and yield can be adequately approximated using a linear formulation. Importantly, this finding aligns with the methodological focus of the thesis. Since the primary objective is not to maximize absolute prediction accuracy, but rather to systematically evaluate how prediction error evolves as forecasts are issued earlier in the growing season, model interpretability and stability are of central importance.

Linear Regression offers several advantages in this context. First, it provides a transparent and consistent mapping between input features and yield, ensuring that observed changes in prediction accuracy can be attributed primarily to data availability and temporal completeness, rather than model-induced variability. Second, linear models exhibit lower sensitivity to noise introduced by temporal filling procedures, making them particularly suitable for assessing early-season predictions where reconstructed pixels are more prevalent.

Moreover, the superior performance of the unregularized linear model suggests that multicollinearity among vegetation indices and weather variables does not critically degrade predictive stability within the selected feature set. This reinforces the suitability of Linear Regression as a robust reference model for subsequent timing-based analyses.

Based on these considerations, Linear Regression is adopted as the final modeling framework for the remainder of this study. Subsequent sections therefore focus on evaluating how prediction accuracy changes as the forecast horizon moves backward from harvest, using this model to isolate the impact of temporal data availability from confounding model complexity effects.

5.5 Prediction Timing vs Accuracy

This section evaluates the central research question of the thesis: how prediction error changes as yield is predicted earlier in the season, and the resulting trade-off between timeliness and accuracy. After completing model selection (Section 5.4), Linear Regression was adopted as the final model due to its superior performance and stability across candidate baselines. The prediction-timing experiment was then conducted by generating yield predictions at multiple forecast dates (March–October) and comparing them against observed yields for each year (2017–2024).

The full results are summarized in Table 4, which reports year-level observed yield (y_{true}), predicted yield (y_{pred}), and error percentage for each month, alongside aggregate performance metrics including R^2 , RMSE, MAE, and error distribution statistics. To support interpretation, month-specific scatter/line plots of actual vs predicted yield are provided and referenced in the discussion as Figures 11–18 (one figure per month). These plots are essential because they reveal not only the average performance but also the structure of errors, such as systematic underestimation/overestimation and sensitivity to extreme years.

Table 4. Table of Prediction Results

PREDICTION MONTH	year	y_true	y_pred	Error %	R2	RMSE	MAE	Max Err	Min Err	Med Err	Avg error
MARCH	2017	87	66.20	23.91	0.136	34.82	30.61	0.74	0.14	0.35	0.42
	2018	71	22.09	68.89							
	2019	95	156.16	64.38							
	2020	72	61.84	14.11							
	2021	66	83.45	26.43							
	2022	52	90.48	74.01							
	2023	69	51.94	24.72							
	2024	72	102.88	42.90							
APRIL	2017	87	74.90	13.90	0.254	18.73	16.29	0.41	0.03	0.24	0.22
	2018	71	68.75	3.17							
	2019	95	67.58	28.87							
	2020	72	42.60	40.83							
	2021	66	78.31	18.64							
	2022	52	35.70	31.34							
	2023	69	62.49	9.43							
	2024	72	96.05	33.41							
MAY	2017	87	44.63	48.70	0.238	24.23	22.64	0.49	0.16	0.32	0.31
	2018	71	50.15	29.37							
	2019	95	110.27	16.07							
	2020	72	45.71	36.51							
	2021	66	84.85	28.56							
	2022	52	73.29	40.94							
	2023	69	57.16	17.16							
	2024	72	47.67	33.79							
JUNE	2017	87	76.79	11.74	0.163	34.45	18.68	1.32	0.07	0.11	0.26
	2018	71	65.21	8.15							
	2019	95	104.48	9.98							
	2020	72	167.01	131.96							
	2021	66	73.34	11.12							
	2022	52	57.28	10.16							
	2023	69	64.20	6.96							
	2024	72	60.49	15.99							

Table 4 (continued)

JULY	2017	87	68.07	21.75	0.544	17.87	15.92	0.44	0.06	0.21	0.22
	2018	71	75.19	5.91							
	2019	95	118.32	24.55							
	2020	72	60.75	15.62							
	2021	66	48.73	26.17							
	2022	52	62.27	19.75							
	2023	69	79.59	15.34							
	2024	72	40.48	43.78							
AUGUST	2017	87	71.76	17.51	0.476	15.80	14.02	0.32	0.06	0.18	0.18
	2018	71	86.68	22.08							
	2019	95	69.16	27.20							
	2020	72	58.34	18.97							
	2021	66	74.73	13.22							
	2022	52	48.75	6.24							
	2023	69	75.66	9.65							
	2024	72	95.12	32.11							
SEPTEMBER	2017	87	90.45	3.97	0.669	16.69	15.12	0.20	0.01	0.07	0.09
	2018	71	73.98	4.19							
	2019	95	89.84	5.44							
	2020	72	70.98	1.42							
	2021	66	75.31	14.10							
	2022	52	57.28	10.16							
	2023	69	55.22	19.97							
	2024	72	78.46	8.97							
OCTOBER	2017	87	79.15	9.03	0.704	7.10	5.86	0.32	0.04	0.15	0.15
	2018	71	58.31	17.87							
	2019	95	105.66	11.22							
	2020	72	59.06	17.98							
	2021	66	78.21	18.51							
	2022	52	68.71	32.13							
	2023	69	74.71	8.27							
	2024	72	74.71	3.76							

A clear seasonal pattern emerges across the forecast months: performance improves as predictions move closer to the end of the season. This trend is consistent with the expected increase in informative signal in vegetation indices and weather accumulations as the crop progresses through its phenological stages and approaches harvest. However, improvements are not perfectly monotonic across all months, indicating that forecast skill depends not only on “how late” the prediction is made, but also on whether the available data capture the most yield-informative stages of growth.

5.5.1 Overall trend: Accuracy improves toward harvest

The aggregated statistics in Table 4 demonstrate that the latest forecast month (October) achieves the strongest overall performance ($R^2 \approx 0.704$, $RMSE \approx 7.10$, $MAE \approx 5.86$). In contrast, early-season months exhibit substantially weaker explanatory power and higher error dispersion. For example, March yields a low R^2 (≈ 0.136) and very high $RMSE/MAE$, reflecting the limited predictive information available at that time.

This finding directly supports the thesis hypothesis: earlier predictions are more uncertain, and the uncertainty decreases as the season progresses. Importantly, the month-by-month plots (Figures 11–18) show that the improvement is not solely a reduction in average error; rather, later months also tend to reduce extreme failures (outlier years), suggesting increasing robustness.

5.5.2 March predictions: Very high uncertainty and unstable year-level errors

March is the earliest forecasting point and exhibits the weakest performance in the table (Table 4) with very large year-to-year deviations and several extreme errors. Year-level errors include:

- Large underestimation (e.g., 2018 predicted ~ 22 vs true 71 $\rightarrow \sim 69\%$ error)
- Large overestimation (e.g., 2019 predicted ~ 156 vs true 95 $\rightarrow \sim 64\%$ error)
- Severe overestimation in 2022 and sizable overestimation in 2024

These behaviors indicate that the model has difficulty distinguishing between years in March using available predictors. At this stage, Sentinel-2 indices often capture early vegetation signals that may not yet reflect final biomass accumulation or stress impacts, while weather effects may be incomplete (insufficient cumulative context).

March predictions exhibit the weakest overall performance among all forecast months. As shown in Table 4, the model achieves a low coefficient of determination ($R^2 \approx 0.136$), accompanied by large RMSE and MAE values. Year-level errors are substantial, with both severe underestimation (e.g., 2018) and strong overestimation (e.g., 2019 and 2022).

This behavior is clearly visible in Figure 11 (March actual vs predicted yield), where data points show wide dispersion and poor alignment with the 1:1 reference line. Predictions in March fail to consistently preserve yield ranking between years, indicating that early-season vegetation and weather signals contain insufficient information to reliably constrain final yield outcomes. March therefore represents a low-confidence prediction stage, suitable only for very coarse early guidance.

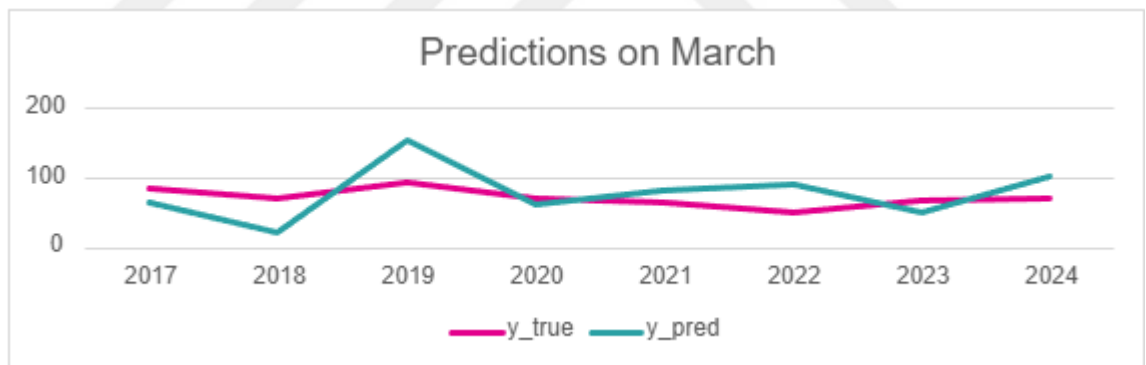


Figure 11. Actual vs Predicted Yield on March

5.5.3 April predictions: Moderate improvement, but still low explanatory power

In April, performance improves relative to March (R^2 rises to ~ 0.254 , RMSE and MAE drop substantially in Table 4). Errors become smaller in several years (e.g., 2018 shows $\sim 3\%$ error), but substantial deviations persist (e.g., 2020 $\sim 41\%$ error, 2024 $\sim 33\%$ error).

This pattern suggests that by April, the feature set begins to capture more meaningful seasonal progression, but the model still struggles in years where the season's

trajectory deviates from typical patterns (e.g., unusual weather anomalies or management differences).

In April, predictive performance improves modestly relative to March. As reported in Table 4, R^2 increases to approximately 0.254, and error magnitudes decline for several years. Nevertheless, notable deviations remain, particularly in years such as 2020 and 2024.

Figure 12 (April actual vs predicted yield) shows a slightly tighter clustering of points compared to March, but the relationship remains weak and scattered. While some years are predicted with reasonable accuracy, others deviate substantially, indicating that April still precedes the most yield-informative phenological stages. As a result, April forecasts remain vulnerable to interannual variability and early-season noise.

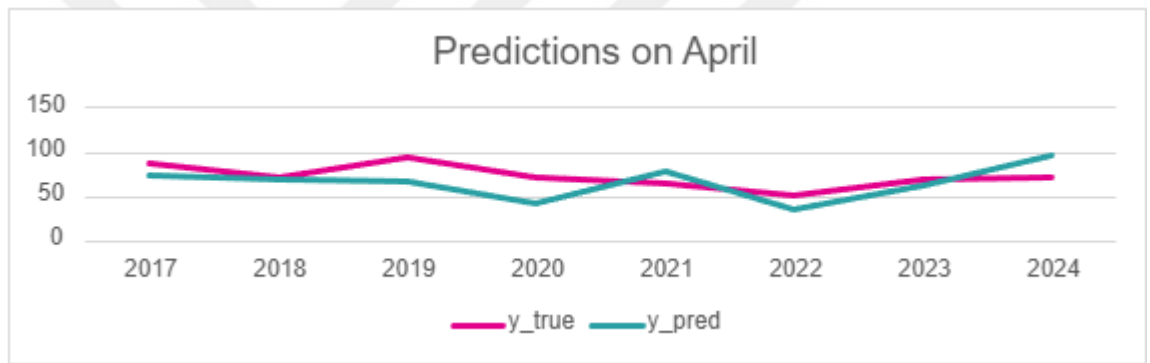


Figure 12. Actual vs Predicted Yield on April

5.5.4 May predictions: Not strictly better than April indicative of phenology and noise effects

May results ($R^2 \approx 0.238$) are slightly weaker than April in terms of R^2 , despite similar RMSE/MAE magnitudes (Table 4). This non-monotonic behavior is important: it indicates that forecast skill does not increase smoothly each month.

A plausible interpretation is that May may coincide with a period where:

- Vegetation indices are increasing rapidly (high temporal variability),
- Cloud masking + temporal filling may still affect measurement consistency,
- And early-season signals may not yet be strongly anchored to final yield.

May predictions do not exhibit a clear improvement over April, despite being temporally closer to harvest. As shown in Table 4, R^2 decreases slightly (≈ 0.238), and several years show large relative errors.

This behavior is evident in Figure 13 (May actual vs predicted yield), where both overestimation and underestimation persist across the yield range. The non-monotonic change in performance suggests that May corresponds to a transitional growth phase during which vegetation indices change rapidly. During such periods, predictive relationships may be more sensitive to noise, cloud masking, and temporal filling effects, leading to unstable model behavior.

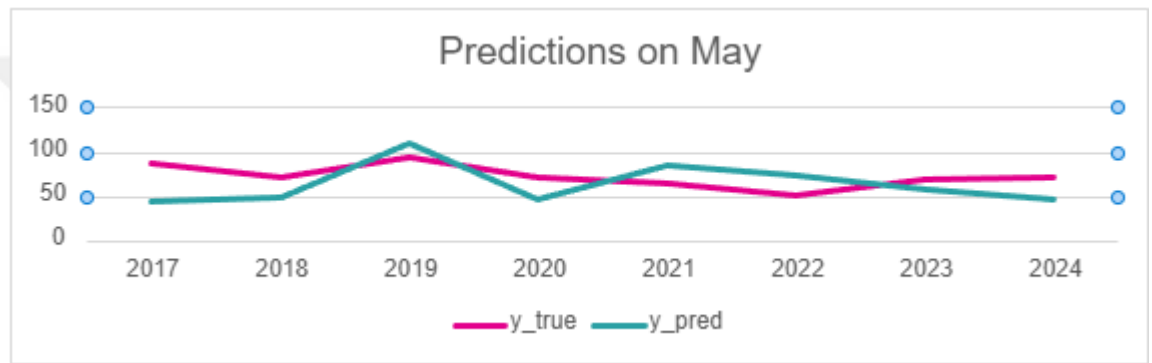


Figure 13. Actual vs Predicted Yield on May

5.5.5 June predictions: Presence of a major outlier year and instability despite mid-season timing

June shows $R^2 \approx 0.163$ with high RMSE (Table 4), which is unexpectedly weak for a mid-season month. The primary driver is clearly visible in the year-level breakdown: 2020 has an extreme overprediction (true 72, predicted ~ 167 ; error $\sim 132\%$). This single year dramatically inflates RMSE and depresses R^2 .

June predictions are characterized by substantial instability despite their mid-season timing. Although some years are predicted reasonably well, performance is dominated by a major outlier in 2020, where yield is strongly overestimated.

This effect is reflected numerically in Table 4, where R^2 remains low (≈ 0.163) and RMSE increases significantly. The impact of the outlier is visually unmistakable in Figure 14 (June actual vs predicted yield), which shows a single point far removed from the main data cluster. This result demonstrates that mid-season predictions can

still be highly sensitive to anomalous conditions or feature distortions, and that temporal proximity alone does not guarantee stability.

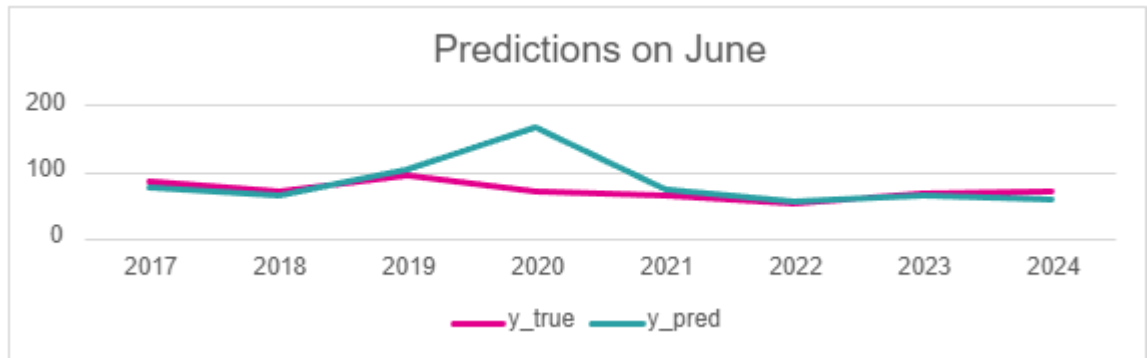


Figure 14. Actual vs Predicted Yield on June

Interpretation-wise, June's instability suggests one (or several) of the following:

1. Atypical seasonal trajectory in 2020 (e.g., weather anomaly, disease, management change) that broke the learned relationship.
2. Feature distortion in a key time window (cloud masking/filling) that produced a misleading vegetation signal.
3. A model sensitivity to specific feature patterns that appear in some years but not others.

5.5.6 July predictions: Major step up in predictive skill

July marks a strong improvement, with R^2 increasing to ~ 0.544 and RMSE/MAE decreasing (Table 4). Although some years still carry substantial error (e.g., 2024 $\sim 44\%$ error), the general relationship between predicted and observed yield becomes notably more stable.

A marked improvement in predictive skill occurs in July. As summarized in Table 4, R^2 rises sharply to approximately 0.544, while RMSE and MAE decrease substantially relative to earlier months.

Figure 15 (July actual vs predicted yield) confirms this improvement, showing tighter clustering of points around the diagonal and improved preservation of yield ranking between years. Although some individual years still exhibit moderate error, July represents the first forecast month where predictions become consistently informative.

This indicates that by mid-summer, accumulated vegetation and meteorological signals capture key determinants of final yield.

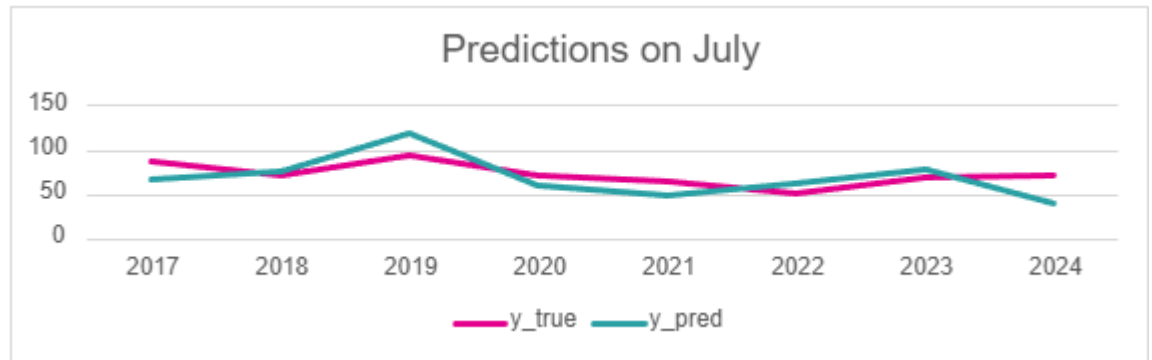


Figure 15. Actual vs Predicted Yield on July

This supports an important applied conclusion: July may represent a practical “decision window” where predictions are early enough to be operationally useful yet accurate enough to be defensible.

5.5.7 August predictions: Best RMSE/MAE among mid-late months, but R^2 slightly lower than July/September

August yields $R^2 \approx 0.476$, RMSE ≈ 15.8 , MAE ≈ 14.0 (Table 4). Compared to July, R^2 is slightly lower, but error levels remain relatively controlled. Several years have modest errors (e.g., 2022 $\sim 6\%$), while others remain large (e.g., 2018 $\sim 22\%$, 2024 $\sim 32\%$).

August predictions maintain relatively low error levels, with RMSE and MAE continuing to decrease, although R^2 (≈ 0.476) is slightly lower than in July. As shown in Table 4, most years exhibit moderate errors, with fewer extreme failures.

In Figure 16 (August actual vs predicted yield), predictions align well with observations for several years, though some systematic deviations remain. These results suggest that August provides a stable prediction window, though interannual variability and late-season effects may still limit full convergence toward final yield.

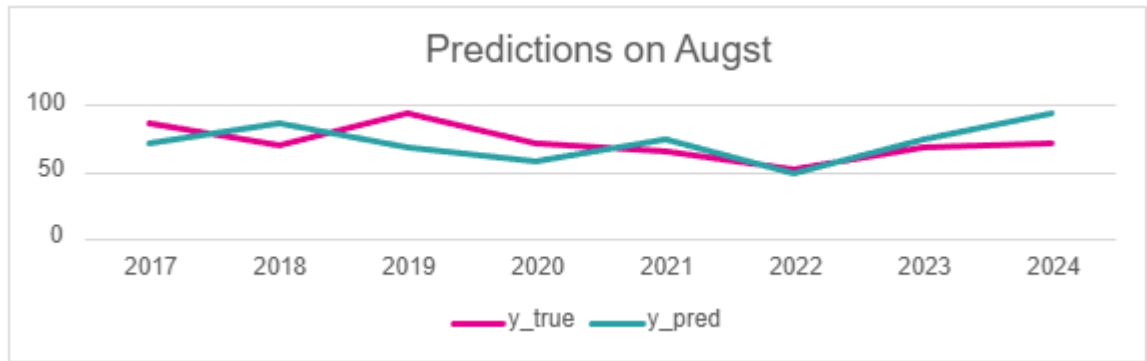


Figure 16. Actual vs Predicted Yield on Augst

5.5.8 September predictions: Peak stability across months before harvest

September achieves one of the strongest fits prior to harvest with $R^2 \approx 0.669$ (Table 4). Interestingly, September's RMSE/MAE are still larger than October's, but overall explanatory power is high, implying strong ranking and alignment between years.

September predictions exhibit one of the strongest performances prior to harvest. As reported in Table 4, R^2 increases to approximately 0.669, indicating strong explanatory power across years.

This behavior is visually supported by Figure 17 (September actual vs predicted yield), where points cluster closely around the diagonal with reduced dispersion. Although some years still show notable deviations, September predictions demonstrate a high degree of robustness and consistency, reflecting the availability of near-complete seasonal information.

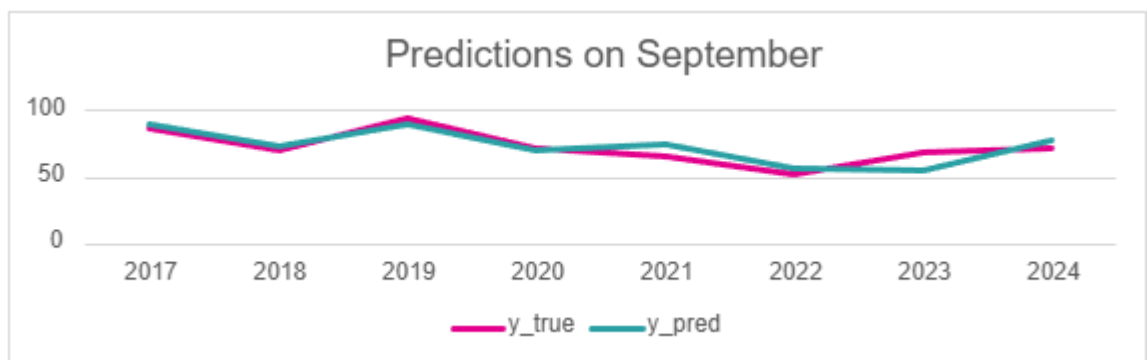


Figure 17. Actual vs Predicted Yield on September

September is often where remote sensing is highly informative because canopy conditions (biomass, chlorophyll, stress signatures) and accumulated weather effects

are strongly linked with final yield outcomes. The results are consistent with that agronomic logic.

5.5.9 October predictions: Best overall performance and “reference point” for the thesis trade-off

October is the final month listed and performs best overall: $R^2 \approx 0.704$, RMSE ≈ 7.10 , MAE ≈ 5.85 (Table 4). Year-level errors are generally low, with several years near or below $\sim 5\%$ error (e.g., 2020 $\sim 1.4\%$, 2017 $\sim 4.0\%$, 2018 $\sim 4.2\%$, 2019 $\sim 5.4\%$). Some deviations remain (e.g., 2023 $\sim 20\%$ error), but the overall behavior is robust.

October predictions achieve the highest overall accuracy among all forecast months. As shown in Table 4, the model reaches $R^2 \approx 0.704$, with the lowest RMSE and MAE values observed across the study.

Figure 18 (October actual vs predicted yield) clearly illustrates tight clustering near the 1:1 line and minimal dispersion relative to earlier months. Most years are predicted with low relative error, confirming October as the high-confidence benchmark representing best-case performance under near-complete seasonal information. This month therefore serves as the reference point against which earlier predictions are evaluated in terms of accuracy loss.

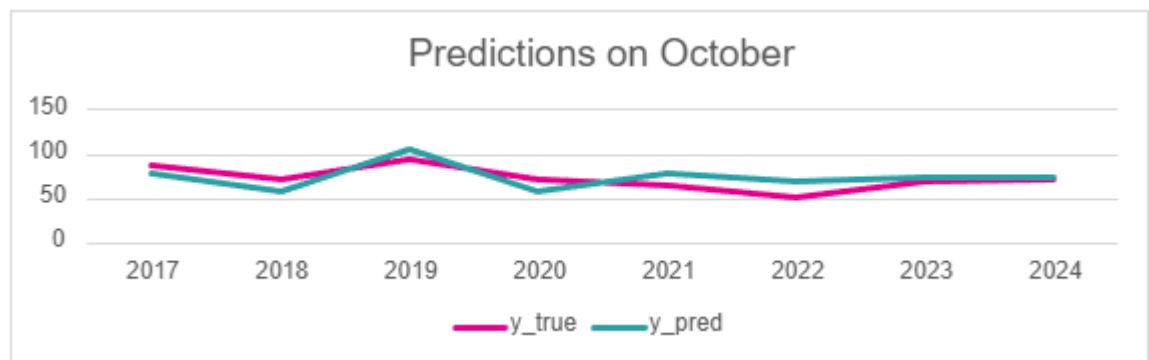


Figure 18. Actual vs Predicted Yield on October

The evolution of prediction performance across forecast months is summarized in Figure 19, which presents the coefficient of determination (R^2) for each prediction timing. A clear upward trend is observed from March to September, with R^2 increasing from approximately 0.13 in March to around 0.704 in October. This pattern quantitatively demonstrates the progressive improvement in model explanatory power

as the growing season advances and more phenological information becomes available.

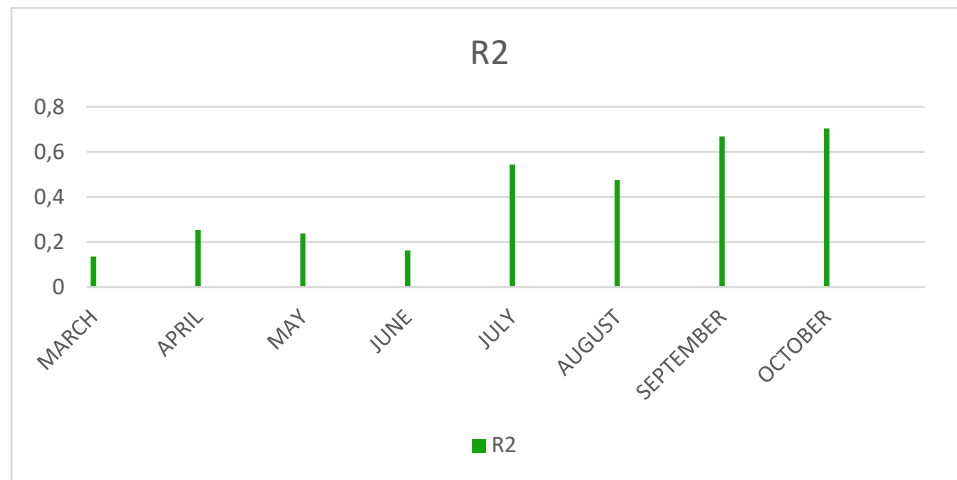


Figure 19. R² Score For Each Month

The progression, however, is not strictly monotonic. A temporary reduction in R^2 is observed in May, followed by a more pronounced improvement in June and July. The May decrease is consistent with the instability discussed in Section 5.5.3 and likely reflects heightened phenological sensitivity during early canopy development. A slight change in performance is also observed in October. Although October exhibits the highest overall R^2 , this period includes a small number of years with moderate overestimation, indicating that even near-harvest predictions retain residual uncertainty despite their strong overall fit.

Overall, Figure 19 provides a concise summary of the central thesis finding: predictions issued earlier in the season exhibit substantially lower explanatory power, while predictive performance improves as additional observational information becomes available. The rate of improvement is particularly pronounced between March and July, after which further gains become more incremental, indicating diminishing returns as the season approaches harvest.

CHAPTER 6: CONCLUSION AND FUTURE WORK

This thesis investigated the trade-off between prediction timing and yield estimation accuracy in remote sensing-based agricultural yield forecasting. Using multi-year Sentinel-2 satellite imagery and meteorological data from the Copernicus Climate Data Store, a systematic modeling framework was developed to evaluate how prediction uncertainty evolves as forecasts are generated progressively earlier in the growing season. Central to this analysis was the integration of cloud masking and temporal filling procedures, enabling the construction of continuous vegetation index time series while preserving data integrity.

The results demonstrate that prediction accuracy is strongly dependent on forecast timing. Early-season predictions, particularly those generated in March and April, exhibit high uncertainty and substantial interannual variability, reflecting limited phenological information and greater sensitivity to reconstructed data. As the season progresses, predictive performance improves markedly, with July and August representing a practical balance between timeliness and accuracy, and September–October providing high-confidence estimates approaching best-case performance. These findings quantitatively confirm the existence of a clear and measurable trade-off between early decision support and predictive reliability.

A key methodological outcome of this study is the identification of Linear Regression as a robust and suitable modeling framework for analyzing prediction timing effects. Despite its simplicity, the linear model outperformed more complex alternatives, including RFs, NNs, and GB variants, in terms of stability and explanatory power. This result highlights that, for temporally aggregated satellite-derived features and limited sample sizes, increased model complexity does not necessarily translate into improved performance. Instead, transparent and stable models are better suited for isolating the effects of data availability and seasonal progression on prediction accuracy.

Importantly, this work contributes beyond model comparison by explicitly linking prediction error behavior to data completeness and seasonal information content. The analysis of temporal filling demonstrated that early-season datasets rely more heavily

on reconstructed pixels, which amplifies uncertainty during periods of rapid vegetation change. As observational coverage increases later in the season, the influence of filled data diminishes, allowing the model to rely more directly on measured signals. This relationship provides a critical interpretative framework for understanding why early predictions are inherently less reliable, regardless of model choice.

While the results indicate that satellite-based yield forecasting is subject to unavoidable early-season limitations, they also underline the value and relevance of this approach. Satellite imagery offers unparalleled spatial coverage, historical depth, and operational feasibility at regional to national scales. The ability to quantify how prediction accuracy evolves over time is itself a significant contribution, as it enables informed decision-making about when forecasts are sufficiently reliable for specific agricultural applications.

At the same time, the findings point toward clear opportunities for future improvement. Yield prediction accuracy could be substantially enhanced through the integration of high-resolution UAV imagery and in-situ sensor networks. UAVs provide finer spatial and temporal resolution, reduced cloud interference, and the ability to capture field-level heterogeneity that is often averaged out in satellite observations. Similarly, local sensor measurements, such as soil moisture, canopy temperature, and microclimatic variables, can supply highly specific information that directly reflects crop conditions within a given field. When combined with accurate, field-level yield tracking, such data sources can significantly reduce uncertainty and enable higher predictive performance.

However, these localized approaches come with trade-offs of their own, including higher operational cost, limited spatial scalability, and increased logistical complexity. In this context, the present study serves as a complementary and foundational contribution. It establishes a benchmark for what can be achieved using widely available satellite and climate data and provides a structured methodology for evaluating prediction timing effects. Rather than competing with UAV or sensor-based systems, this work defines the performance envelope within which satellite-driven forecasting operates and clarifies where localized data collection can offer the greatest marginal benefit.

In conclusion, this thesis demonstrates that yield prediction accuracy is fundamentally constrained by the timing and completeness of available information. By systematically quantifying this constraint, the study advances understanding of early-season yield forecasting and provides practical guidance on the reliability of predictions at different stages of the growing season. Future research integrating satellite data with UAV observations and local sensor networks holds strong potential to further improve accuracy, but the methods and insights developed here remain essential for scalable, data-driven agricultural monitoring and decision support.



REFERENCES

- ESA, “Sentinel-2 User Handbook,” European Space Agency, 2015.
- J. Drusch et al., “Sentinel-2: ESA’s optical high-resolution mission for GMES operational services,” *Remote Sensing of Environment*, vol. 120, pp. 25–36, 2012.
- J. W. Rouse et al., “Monitoring vegetation systems in the Great Plains with ERTS,” *NASA Special Publication*, vol. 351, pp. 309–317, 1974.
- C. Tucker, “Red and photographic infrared linear combinations for monitoring vegetation,” *Remote Sensing of Environment*, vol. 8, pp. 127–150, 1979.
- B. Wardlow and S. Egbert, “Large-area crop mapping using time-series MODIS 250 m NDVI data,” *Remote Sensing of Environment*, vol. 112, pp. 1096–1116, 2008.
- A. A. Gitelson and M. Merzlyak, “Signature analysis of leaf reflectance spectra: algorithm development for remote sensing of chlorophyll,” *Journal of Plant Physiology*, vol. 148, pp. 494–500, 1996.
- D. Haboudane et al., “Hyperspectral vegetation indices and novel algorithms for predicting green LAI of crop canopies,” *Remote Sensing of Environment*, vol. 90, pp. 337–352, 2004.
- A. Gitelson, “Wide dynamic range vegetation index for remote quantification of crop biophysical characteristics,” *Journal of Plant Physiology*, 2004.
- A. A. Gitelson et al., “Relationships between leaf chlorophyll content and spectral reflectance and algorithms for non-destructive chlorophyll assessment in higher plant leaves,” *Journal of Plant Physiology*, 2003.
- D. Dash and P. Curran, “The MERIS terrestrial chlorophyll index,” *International Journal of Remote Sensing*, vol. 25, pp. 5403–5413, 2004.
- L. Breiman, “Random forests,” *Machine Learning*, vol. 45, pp. 5–32, 2001.
- F. Johnson, “Maize yield prediction using random forest regression,” *Agricultural Systems*, 2016.
- A. Jeong et al., “Random forests for global crop yield modeling,” *PLOS ONE*, 2016.
- J. Friedman, “Greedy function approximation: A gradient boosting machine,” *Annals of Statistics*, 2001.
- S. Khaki and L. Wang, “Crop yield prediction using deep neural networks,” *Frontiers in Plant Science*, 2019.

- S.Y. Han, T.F.A. Bishop, P. Filippi, Data-driven, early-season forecasts of block sugarcane yield for precision agriculture, *Field Crops Research* 276 (2022) 108360.
- T. Hastie, R. Tibshirani, J. Friedman, *The Elements of Statistical Learning*, Springer, 2009 (Chapter 7: Model Assessment and Selection).
- B.C. Gao, NDWI-A normalized difference water index for remote sensing of vegetation liquid water from space, *Remote Sensing of Environment*, 58 (1996), 257–266.
- A. Pandey, M. Patel, R. Parida, Crop moisture stress detection using NDMI and other spectral indices, *International Journal of Remote Sensing*, 40 (2019), 450–470.
- Liu, J., Pattey, E., & Jégo, G. (2020). Assessment of vegetation indices for regional crop syield prediction using MODIS data. *Agricultural and Forest Meteorology*, 291, 108096.
- Potgieter, A. B., et al. (2016). Multispectral remote sensing for wheat yield prediction. *Remote Sensing of Environment*, 174, 339–350.
- Peng, Y., et al. (2018). Assessment of crop nitrogen status using red-edge vegetation indices. *Field Crops Research*, 201, 135–143.
- Bolton, D. K., & Friedl, M. A. (2013). Forecasting crop yield using time-series vegetation indices. *Remote Sensing of Environment*, 129, 1–13.
- Abbas, A., van der Ploeg, M., & van der Ploeg, M. J. (2021). *Crop Yield Estimation and Interpretability With Gaussian Processes*.
- Abbas, A., Sishodia, R. P., & Ray, R. L. (2020). *Crop Yield Prediction through Proximal Sensing and Machine Learning Algorithms*.
- Arshad, S., Bakhsh, A., Ali, M., Mehmood, H., Ali, M., Nafees, M., & Qureshi, M. K. (2023). *Applicability of machine learning techniques in predicting wheat yield based on remote sensing and climate data*.
- Chlingaryan, A., Sukkarieh, S., & Whelan, B. (2018). *Machine learning approaches for crop yield prediction and nitrogen status estimation in precision agriculture*. Computers and Electronics in Agriculture, 151, 61–69.

- Filippi, P., Jones, E., Whelan, B., & Bishop, T. (2019). *An approach to forecast grain crop yield using multilayered, multifarm data sets and machine learning*. Precision Agriculture, 20(1), 101–120.
- Hoffman, A. L., Kemanian, A. R., & Forest, C. E. (2018). *Analysis of climate signals in the crop yield record of sub-Saharan Africa*. Global Change Biology, 24(1), 143–157.
- Kim, N., Ha, J., & Kim, S. (2019). *A Comparison Between Major Artificial Intelligence Models for Crop Yield Prediction: Case Study of the Midwestern United States, 2006–2015*.
- Leo, S., King, G., & Mallawaarachchi, T. (2020). *Predicting within-field cotton yields using publicly available datasets and machine learning*. Computers and Electronics in Agriculture, 179, 105826.
- Li, T., Zhang, X., & Wei, X. (2020). *Yield prediction model of rice and wheat crops based on ecological distance algorithm*.
- Paudel, D., Dreccer, M. F., & Webber, H. (2022). *Machine learning for regional crop yield forecasting in Europe*. Agricultural and Forest Meteorology, 322, 108940.
- Paudel, D., & co-authors. (2021). *Machine learning for large-scale crop yield forecasting*.
- Seireg, H., & Wang, J. (2022). *Ensemble Machine Learning Techniques Using Computer Simulation Data for Wild Blueberry Yield Prediction*.
- Tian, L., Li, J., Li, X., & Wang, L. (2020). *Yield prediction model of rice and wheat crops based on ecological distance algorithm*.
- Torsoni, G. B., de Oliveira, G., & Marin, F. (2022). *Soybean yield prediction by machine learning and climate*.
- Van Klompenburg, T., Kassahun, A., & Catal, C. (2020). *Crop yield prediction using machine learning: A systematic literature review*. Computers and Electronics in Agriculture, 178, 105709.
- Xu, H., Liu, Y., & co-authors. (2021). *Machine learning approaches can reduce environmental data requirements for regional yield potential simulation*.

Zhao, G., Webber, H., & Wolf, J. (2015). *Demand for multi-scale weather data for regional crop modeling*. *Agricultural and Forest Meteorology*, 200, 156–165.

